AWARD NUMBER: W81XWH-13-1-0284

TITLE: **Biological and Clinical Characterization of Novel lncRNAs Associated with Metastatic Prostate Cancer**

PRINCIPAL INVESTIGATOR: Rohit Malik, Ph.D.

CONTRACTING ORGANIZATION: Regents of the University of Michigan
Ann Arbor, Michigan 48109-1274

REPORT DATE: November 2015

TYPE OF REPORT: Final

PREPARED FOR: U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

DISTRIBUTION STATEMENT: Approved for public release; distribution is unlimited.

The views, opinions and/or findings contained in this report are those of the author(s) and should not be construed as an official Department of the Army position, policy or decision unless so designated by other documentation.

REPORT DOCUMENTATION PAGE	*Form Approved* *OMB No. 0704-0188*

1. REPORT DATE November 2015	**2. REPORT TYPE** Final	**3. DATES COVERED** 15Aug2013 - 14Aug2015

4. TITLE AND SUBTITLE Biological and Clinical Characterization of Novel lncRNAs Associated with Metastatic Prostate Cancer	**5a. CONTRACT NUMBER** **5b. GRANT NUMBER** W81XWH-13-1-0284 **5c. PROGRAM ELEMENT NUMBER**
6. AUTHOR(S) Rohit Malik, Ph.D. E-Mail: romalik@med.umich.edu	**5d. PROJECT NUMBER** **5e. TASK NUMBER** **5f. WORK UNIT NUMBER**
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) Regents of the University of Michigan 3003 S. State Street Ann Arbor, MI 48109-1274	**8. PERFORMING ORGANIZATION REPORT NUMBER**
9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES) U.S. Army Medical Research and Materiel Command Fort Detrick, Maryland 21702-5012	**10. SPONSOR/MONITOR'S ACRONYM(S)** **11. SPONSOR/MONITOR'S REPORT NUMBER(S)**

12. DISTRIBUTION / AVAILABILITY STATEMENT

Approved for Public Release; Distribution Unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT

Despite improvements in medical treatments over the past three decades, prostate cancer remains the second most common cause of cancer related deaths among U.S. men. According to National Cancer Institute, in 2012 ~241,000 prostate cancer diagnosis and ~28,000 related deaths of American men were estimated. Androgen deprivation, surgery, and/or radiotherapy in combination with chemotherapy has proven to be effective in treating patients that display localized disease; however, progression to hormone refractory aggressive disease in a subset of prostate cancer patients remains the primary cause of mortality. The molecular mechanisms that contribute to the progression of localized disease into an aggressive disease remain largely unknown. Long non-coding RNAs (lncRNAs) have recently emerged as key players in tumor biology and can potentially serve as promising biomarkers. However, the vast majority of lncRNAs remain undiscovered or uncharacterized entities and their roles in prostate cancer are unclear.

15. SUBJECT TERMS

Nothing listed

16. SECURITY CLASSIFICATION OF:			**17. LIMITATION OF ABSTRACT**	**18. NUMBER OF PAGES**	**19a. NAME OF RESPONSIBLE PERSON** USAMRMC
a. REPORT Unclassified	**b. ABSTRACT** Unclassified	**c. THIS PAGE** Unclassified	UU	60	**19b. TELEPHONE NUMBER** *(include area code)*

Table of Contents

1. INTRODUCTION:

Despite improvements in medical treatments over the past three decades, prostate cancer remains the second most common cause of cancer related deaths among U.S. men. According to National Cancer Institute, in 2012 ~241,000 prostate cancer diagnosis and ~28,000 related deaths of American men were estimated. Androgen deprivation, surgery, and/or radiotherapy in combination with chemotherapy has proven to be effective in treating patients that display localized disease; however, progression to hormone refractory aggressive disease in a subset of prostate cancer patients remains the primary cause of mortality. The molecular mechanisms that contribute to the progression of localized disease into an aggressive disease remain largely unknown. Long non-coding RNAs (lncRNAs) have recently emerged as key players in tumor biology and can potentially serve as promising biomarkers. However, the vast majority of lncRNAs remain undiscovered or uncharacterized entities and their roles in prostate cancer are unclear. Recent advances by our lab using high-throughput sequencing of prostate cancer tissues have led to the systematic identification of lncRNAs aberrantly expressed in prostate cancer that may play a role in prostate cancer progression. The two main objectives of this proposal were: 1) To evaluate the role of lncRNAs in the progression of prostate cancer and 2) To explore the potential of prostate-specific lncRNAs to reliably predict disease progression.

2. KEYWORDS:

Prostate Cancer, Long non-coding RNA, Androgen receptor, PCAT, PRCAT, DHT, RNA, TCGA, SChLAP-1

3. OVERALL PROJECT SUMMARY:

The overall goal of this project was to characterize long non-coding RNAs that may play an important role in the progression of prostate cancer. We hypothesized that lncRNAs may contribute to the development of aggressive disease and can be used as biomarkers of disease progression. To test our hypothesis we proposed the following three specific aims:

Specific Aim-1: To characterize nominated lncRNAs associated with metastatic castration-resistant prostate cancer.

Specific Aim 2: To elucidate the function of lncRNA in metastatic castrate resistant prostate cancer

Specific Aim 3: To identify lncRNAs that serve as potential biomarkers of disease progression.

<u>Summary of results</u>

Using RNA sequencing data from prostate cancer tissue and benign prostate gland, we were able to identify several lncRNAs that were differentially expressed in prostate cancer. As part of this proposal, we specifically focused on characterizing two such lncRNAs PCAT29 and PCAT51/PRCAT47.

PCAT29

PCAT29 (Prostate cancer associated transcript-29) was identified from differential expression analysis as a lncRNA overexpressed in prostate cancer. We further characterized PCAT29 and show that AR directly binds to the promoter of PCAT29 and suppress its expression. Functionally, PCAT29 acts as a tumor suppressor as its knockdown increased proliferation and its overexpression suppressed growth and metastases of prostate tumors cells. Furthermore, PCAT29 expression was able to predict for poor outcomes in prostate cancer patients. Taken together, our data show that PCAT29 is an androgen-regulated tumor suppressor in prostate cancer. A manuscript describing these findings was published earlier [1] (see appendix).

PRCAT47

Regulation of PCAT29 by androgen receptor (AR) prompted us to systematically study the effects of AR on regulation of lncRNAs. To address this, we performed RNA-Seq on prostate cancer cell lines stimulated with androgen (AR agonist) and performed differential expression analysis to identified lncRNAs regulated by AR. To identify AR-regulated lncRNAs that potentially play important roles in prostate cancer, we calculated the expression of all the transcripts in prostate cancer tissue RNA-seq data obtained from the consortia such as The Cancer Genome Atlas (TCGA), Stand-Up to Cancer (SU2C) [2] as well as those generated in our laboratory [3]. We then combined the two analyses to identify genes that were regulated by AR and were differentially expressed in primary and metastatic prostate cancer. We further utilized our recently queried data set of lncRNAs from multiple cancer type [4] to focus on lncRNAs that displayed both prostate cancer- and prostate tissue-specific expression that were also expressed at high levels (> 10 FPKM). From both the above analysis combined, we nominated a lncRNA called PCAT51/PRCAT47 as one of the top-ranking genes that is AR-regulated and differentially expressed in primary and metastatic prostate cancer compared to normal tissue. We show that PRCAT47 is important for cell survival by regulating AR signaling. A manuscript describing these results is in preparation and will be submitted soon.

SChLAP-1

SChLAP-1 (second chromosome locus associated with prostate-1) was previously identified in our laboratory and was show to be overexpressed in a subset of metastatic prostate cancer patients. In these patients, SChLAP1 promotes formation of lethal prostate cancer by regulating the function of SWI/SNF complex. Furthermore, we also show that SChLAP-1 is an excellent prognostic biomarker as its expression can predict for poor outcomes in prostate cancer patients. A manuscript describing these results is already published [5] (see appendix).

Additional objectives/projects accomplished during the fellowship term.

Landscape of lncRNA in cancers

In addition to the above mentioned aims and objectives, we also defined the landscape of long non-coding RNA in cancers (including prostate cancer). Using RNA-seq data from 18 different cancer types (approximately 6800 samples) we identified 7,942 lineage- or cancer-associated lncRNA. The lncRNA landscape will be important to identify novel lncRNAs that may play role in cancer pathogenesis and will be valuable for the discovery of novel biomarkers of disease progression. A manuscript describing these results is already published [4] (see appendix).

The following section will highlight the progress made in each subaim/task proposed in the grant.

Specific Aim-1: To characterize nominated lncRNAs associated with metastatic castration-resistant prostate cancer. (Month 1-18)

We hypothesize that lncRNAs may contribute to the development of aggressive disease and can be used as biomarkers of disease progression. To test our hypothesis we proposed to identify and systematically characterize novel long non-coding RNA that may play important role in prostate cancer progression.

Using RNA-seq on prostate cancer cell lines and tumor tissue, we previously identified several lncRNAs that were upregulated in prostate cancer compared to benign prostate tissues. This aim specifically focused on determining the expression (Task-1), exon structure (Task-2) and *in-vivo* and *in-vitro* function (Task-3-9) of two candidate lncRNAs, PCAT29 and PRCAT47.

PCAT29

Using RNA-seq data from prostate cancer tissues and cell lines, we previously identified around 121 lncRNAs that were differentially expressed in prostate cancer compared to normal [6]. One of the top outlier candidates in this analysis was a lncRNA called prostate cancer-associated transcripts-29 (PCAT29). To understand the function and mechanism of PCAT29, we characterized this lncRNA in greater details. Using the RNA-seq predicted transcript structures; we designed exon spanning primers and assessed the expression of PCAT29 in various prostate cancer cell lines. As shown in **Figure 1**, PCAT29 expression was highest in three AR positive cell lines (VCaP, LNCaP and 22Rv1) and low in AR-negative (DU145 and PC3) and benign cells (PREC and RWPE).

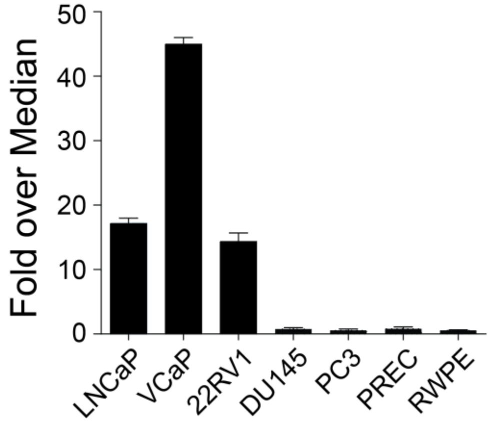


Figure-1: Expression of PCAT29 in various prostate cancer and benign cell lines as determined by qPCR.

We next performed rapid amplification of cDNA ends (RACE) to determine the full exon structure of PCAT29 using LNCaP and VCaP cells and show that it is a 694bp poly-adenylated transcript present on chr15q(23) (**Figure 2**). Furthermore, PCAT29 gene possessed all the chromatin marks (H3K4me3, and DNaseH) that predicts for open chromatin state and are commonly found near or at the transcription start sites (TSSs) (4), suggesting that PCAT29 is an actively transcribed gene (**Figure 2**).

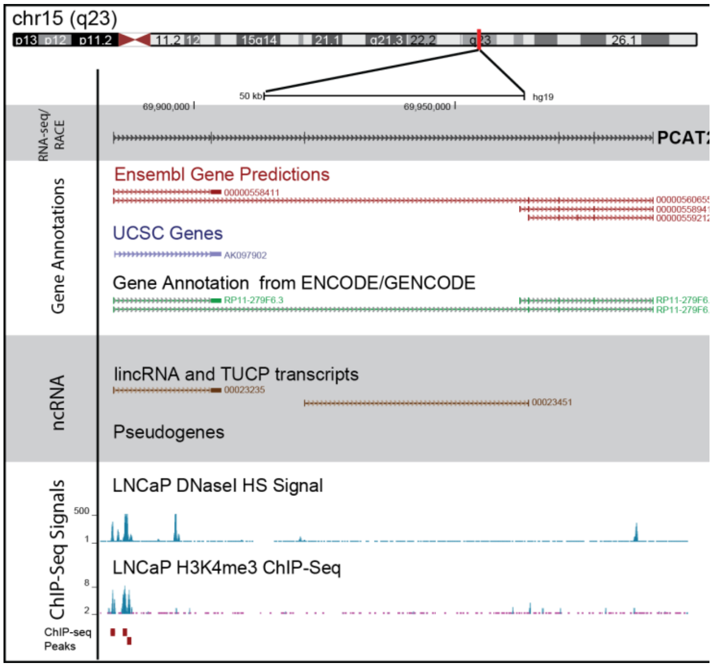


Figure-2: Genome browser view of PCAt29. Exon structure determined by RACE as well as other gene annotation is shown.

Our RNA-seq pipeline identified several lncRNAs that were upregulated in prostate cancer. We validated several of these identified lncRNAs using qPCRs in prostate cancer cell lines and compared the expression in normal prostate cell lines(**Figure-1**). Some of the novel lncRNAs validated are: PCAT29, SChLAP-1, G5303 and PRCAT47. Expression of PCAT29 and SChLAP-1 has been published (see appendix)

Androgen receptor (AR) has been shown to play an important role in the progression of prostate cancer. AR has been shown to regulate variety of genes including lncRNAs. Since PCAT29 expression was higher in AR-positive cell lines, we next examined its regulation by AR. Using previously published AR ChIP-Seq data, we were able to identify an AR peak in the promoter of PCAT29. Upon closer look we identified a canonical AR binding site suggesting that PCAT29 could be regulated by AR (**Figure 3A**). To validate this we performed ChIP-PCR using AR antibody in VCaP cell line and show that indeed AR can bind to the promoter of PCAT29 (**Figure 3B).** To identify the nature of this regulation, we performed qPCR on two AR positive cell lines stimulated with AR, and show that PCAT29 expression is suppressed upon AR stimulation (**Figure 3C)**. This was consistent with the observation that treatment with AR antagonist such as MDV3100 and bicalutamide induces PCAT29 expression (**Figure 3D)**. Furthermore, expression of PCAT29 was induced in LNCaP xenografts 5 days after physical castration (**Figure 3E**). Taken together our data suggests that AR directly binds to the promoter of PCAT29 and suppresses its expression.

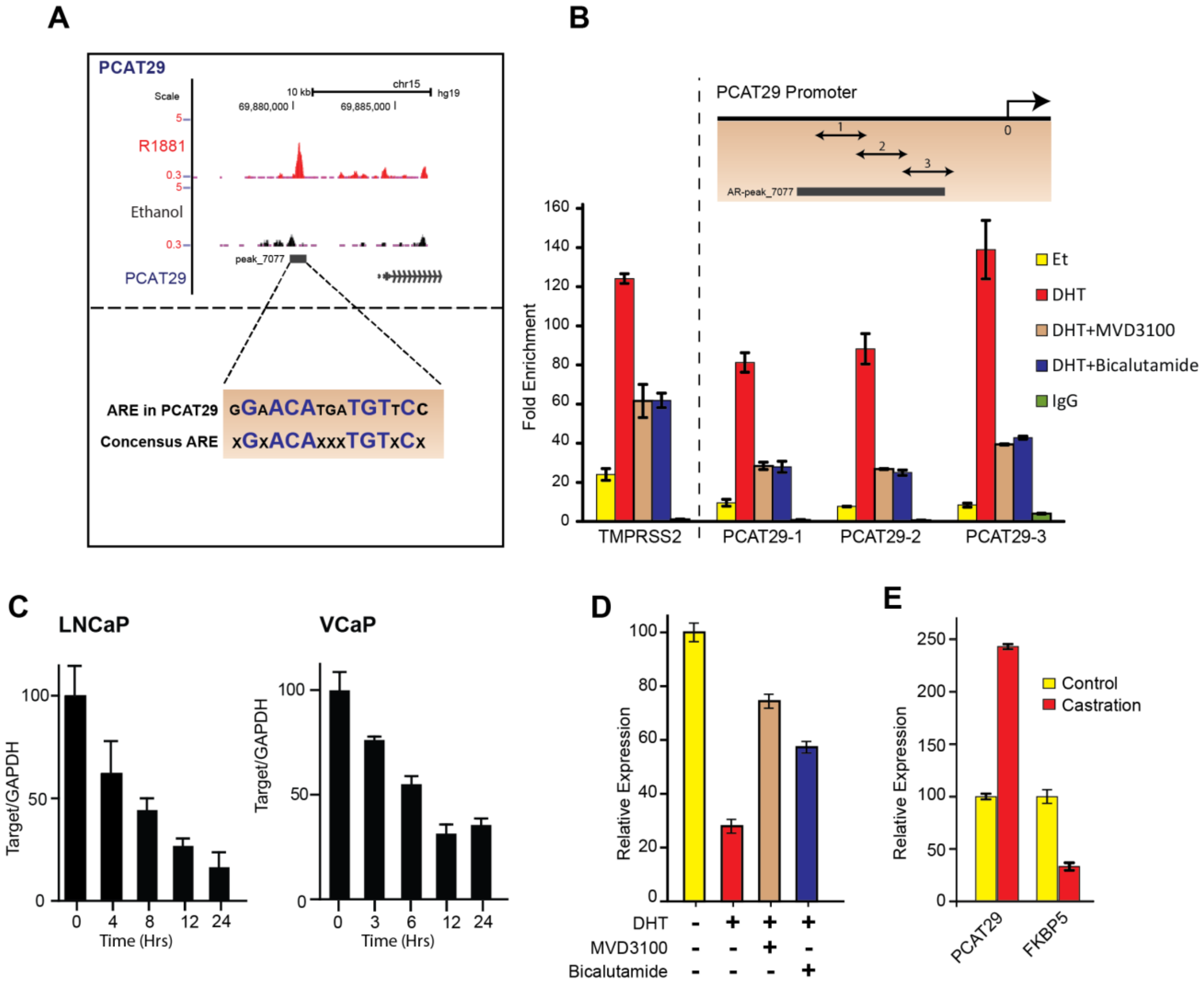


Figure 3 : A. genome browser representation of androgen receptor (AR) binding on the promoter of PCAT29 before and after stimulation with 1 nmol/L R1881. Consensus androgen-responsive elements (ARE) and ARE present in the PCAT29 promoter are shown. **B**. ChIP-PCR to confirm AR occupancy on TMPRSS2 and PCAT29 gene promoter. The y-axis represents AR ChIP enrichment in VCaP cells treated with 10 nmol/L DHT normalized to ethanol (Ethl)-treated cells. **C-D.** Expression of PCAT29 in LNCaP and VCaP cells treated with 10 nmol/L DHT for indicated time points. **E**, Expression of PCAT29 in LNCaP cells treated with 10 nmol/L DHT in the presence or absence of 10 µmol/L MDV3100 or bicalutamide. **F**. Expression of PCAT29 and FKBP5 in LNCaP xenografts obtained from control mice and mice that were physically castrated for 5 days. D,

To investigate the function of PCAT29, we designed two independent shRNAs that can knockdown PCAT29 expression in VCaP and LNCaP cells. Both the shRNAs were able to knockdown PCAT29 to about 70-80% (**Figure 4**).

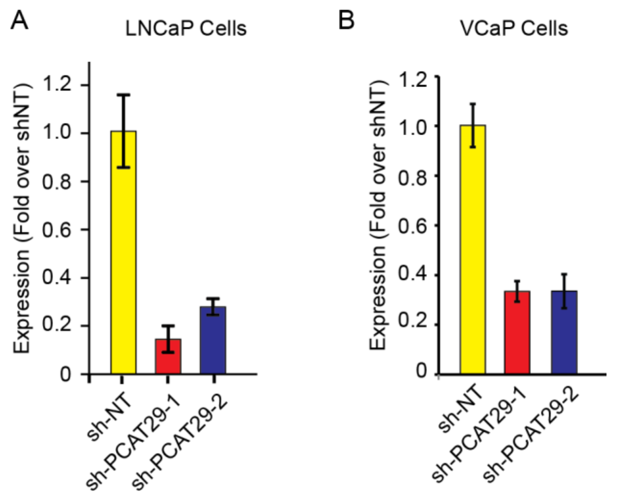


Figure 4: Expression of PCAT29 in LNCaP (A) and VCaP (B) cells.

To assess the function of PCAT29, we performed microarray analysis on cells that were transfected with shRNA targeting PCAT29. Microarray data analysis predicted that PCAT29 may be involved in cell proliferation and migration (see Aim-2 for more details).

To validate this observation, we performed *in-vitro* proliferation and migration assays in LNCaP cells transfected with shRNA targeting PCAT29. As shown in **Figure 5A and B**, knockdown of PCAT29 increased cell proliferation and migration of LNCaP cells, suggesting a tumor suppressive role for PCAT29.

To further validate this observation, we overexpressed full length PCAT29 in a cell line (DU145) that does not express endogenous PCAT29. Upon overexpression, we observed decrease in both proliferation and migration of DU145 cells compared to empty-vector control cells (**Figure 5C-D**).

We next assessed whether similar effects of PCAT29 overexpression could be achieved in vivo. 22RV1 prostate cancer cells overexpressing PCAT29 were implanted on the chick chorioallantoic membrane (CAM) of a chicken egg. Compared to control cells, overexpression of PCAT29 significantly decreased the growth of tumor on the CAM as well as decreased liver metastases (**Figure 6A, B**).

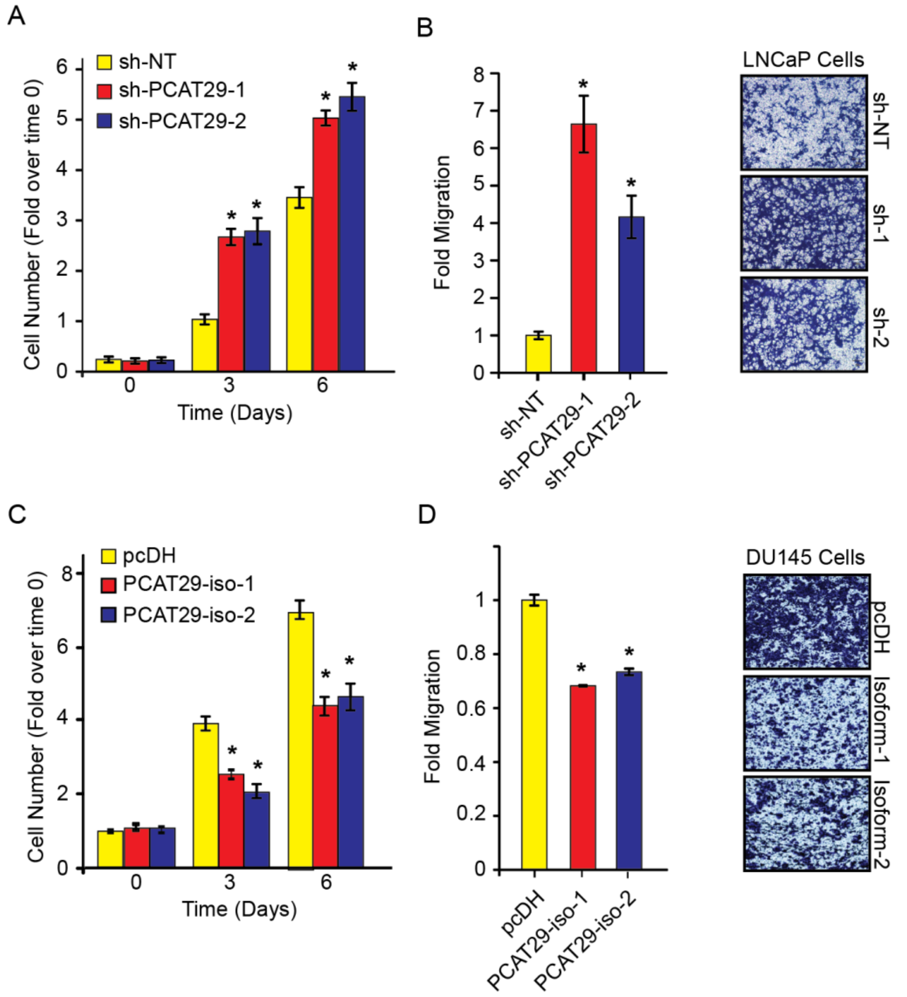


Figure 5: Proliferation and migration of LNCaP cells stably expressing PCAT29 shRNA (A-B) and DU145 cells expressing PCAT29 expression constructs (C-D). Representative micrographs of crystal violet–stained migrated cells are shown as insets. Data, mean ± SEM. *, P < 0.05 by the Student t test.

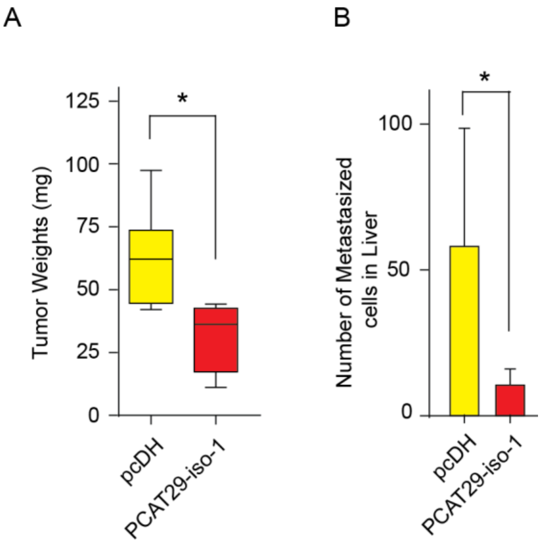
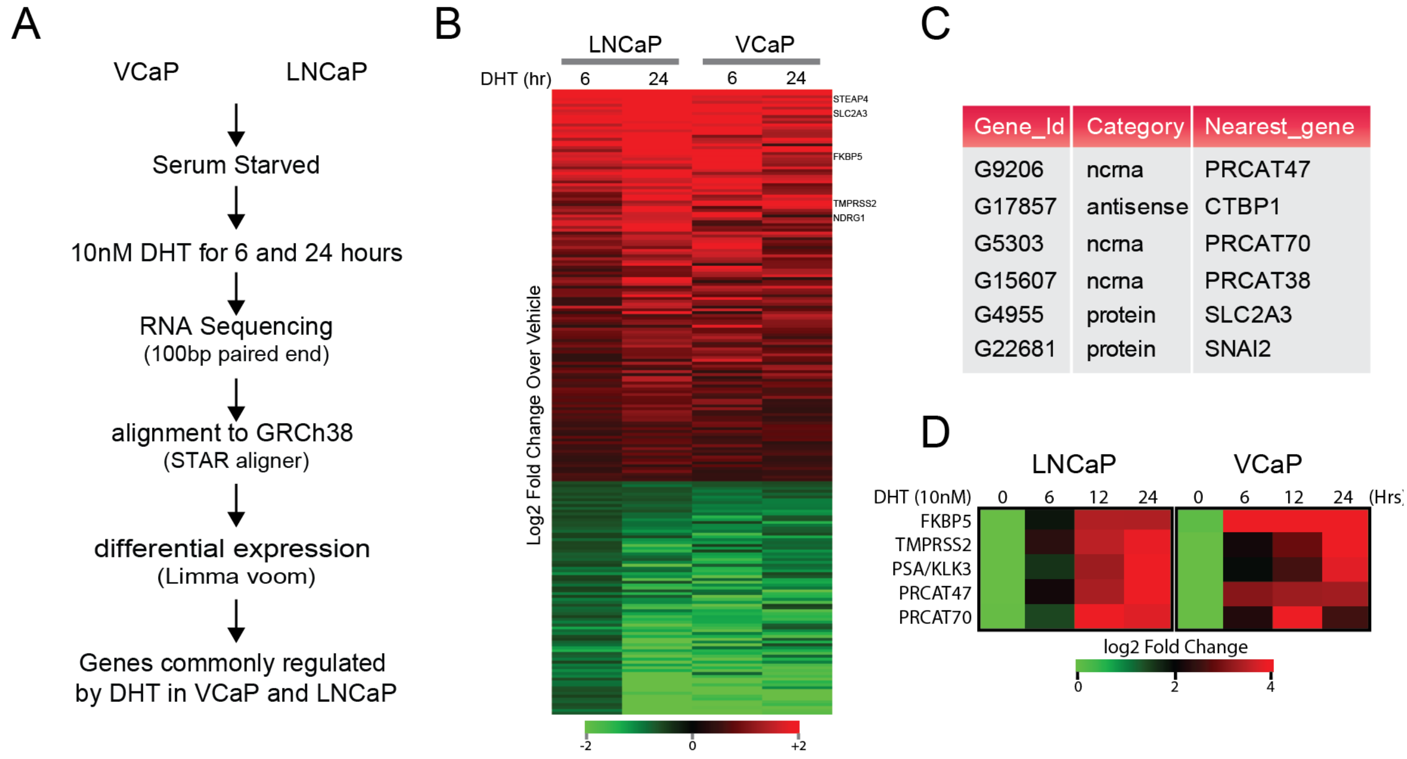


Figure 6: Quantification of tumor weight and metastasis to liver for 22Rv1 cells expressing full length PCAT29 or empty vector (pcDH) in the chick chorioallantoic membrane (CAM) assay. Data are represented as mean ± S.E.M. An asterisk (*) indicated p < 0.05 by Student's T-test.

PRCAT47

Motivated by the observation that PCAT29 was regulated by AR, we decided to perform an unbiased discovery of androgen receptor (AR) regulated long non-coding RNAs in prostate cancer. Transcriptome sequencing (RNA-seq) was performed on AR dependent cell lines VCaP and LNCaP stimulated with DHT for 6 and 24



Gene_Id	Category	Nearest_gene
G9206	ncrna	PRCAT47
G17857	antisense	CTBP1
G5303	ncrna	PRCAT70
G15607	ncrna	PRCAT38
G4955	protein	SLC2A3
G22681	protein	SNAI2

Figure 7: **A**. Schematic representation of experimental outline. **B**. Heatmap of gene expression changes after DHT stimulation in LNCaP and VCaP cells. **C**. List of top 5 candidate gene identified fro mthe analysis. **D.** Validation of 5 coding and non-coding transcripts by q-RT-PCR.

hours (**Figure 7A**). We performed de-novo gene assembly to identify protein coding and non-coding genes that were stimulated with AR **(Figure 7B)**. We further utilized our recently queried data set of lncRNAs from multiple cancer types [4] to focus on highly expressed lncRNAs (>10 FPKM) that displayed both prostate cancer- and tissue-specific expression patterns (**Figure 7C**). From both the above analysis combined, we nominated a lncRNA called PRCAT47 as one of the top-ranking genes that is AR-regulated and differentially expressed in primary and metastatic prostate cancer compared to normal tissue. We also validated the AR regulation aspect of these genes by qPCR in DHT stimulated LNCaP and VCaP cells (**Figure 4D**).

We first performed qPCR to determine expression levels of PRCAT47 in various prostate cell lines. Expression of PRCAT47 was highest in MDA PCa-2b and LNCaP cells **(Figure 8)**. This was further corroborated using RNA-Seq data (**Figure 8 inset**). We then performed random amplification of cDNA ends (RACE) to determine the gene structure of PRCAT47. We show that PRCAT47 is located on chromosome 16 and encodes a 4-exon 2.7kb transcript (**Figure 9**). The length of PRCAT47 was further confirmed in MDA PCa-2b cells by northern blot analysis **(Figure 9 inset).**

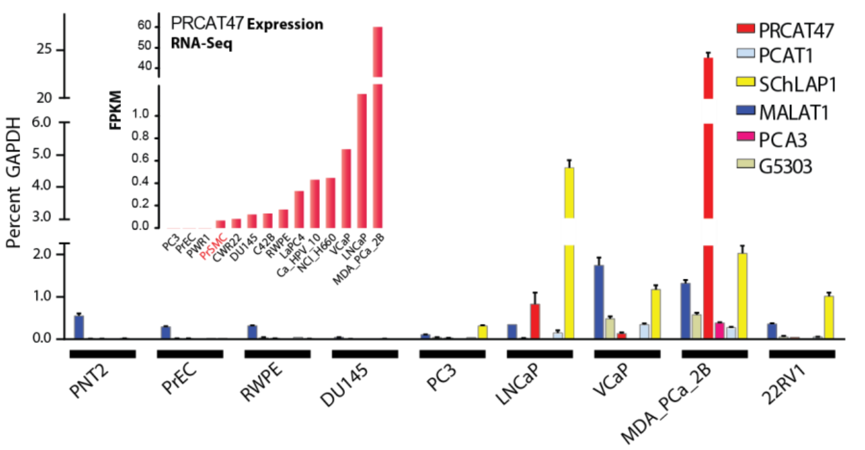


Figure 8: Expression of each lncRNA in prostate cancer cell lines. Inset shows the expression of PCAT51 determined by RNA-seq in prostate cancer cell lines.

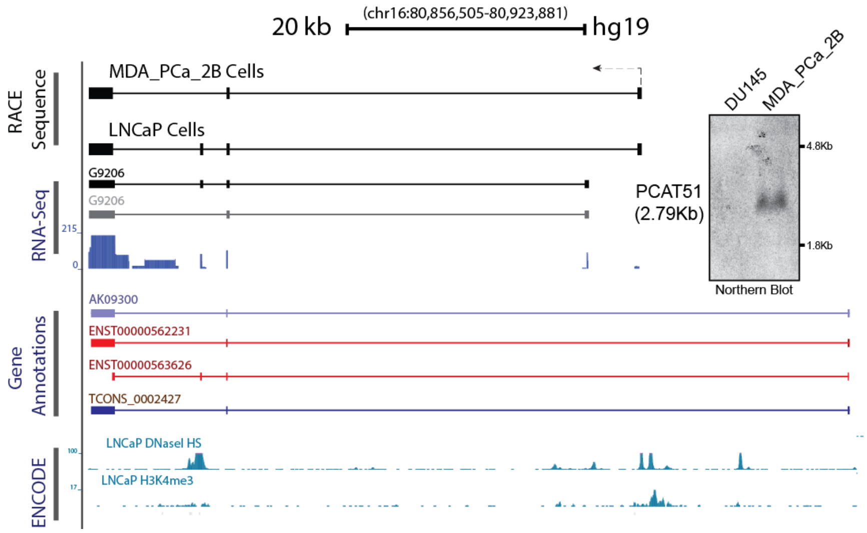


Figure 9: Genome browser view of PRCAT47. Exon structure determined by RACE as well as other gene annotation is shown. Inset shows PRCAT47 northern blot.

To understand the function of PRCAT47, we designed siRNA targeting both PRCAT47 and tested their efficiency of knockdown in cell lines where the expression of lncRNA was high (LNCaP and MDA_PCa_2b cells). In both the cell lines we got 80% knockdown of the gene (**Figure-10A, B**). Gene expression was quantified by qRT-PCR.

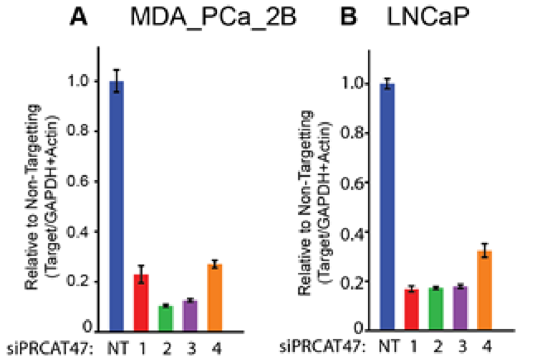


Figure 10: MDA_PCa_2B and LNCaP cells were transfected with 4 independent siRNAs targeting PRCAT47. Expression of PRCAT47 was determined by q-RT-PCR.

Similar to PCAT29, we performed cell proliferation assay using the IncuCyte live cell monitoring to look at the effects of PRCAT47 knockdown. As expected, knockdown of PRCAT47 had a significant effect on the proliferation of MDA PCa-2b and LNCaP cells (**Figure 11A**). No such effect was seen in DU145 cells, an AR-negative cell line that does not express PRCAT47. Knockdown of PRCAT47 also led to an increase in apoptosis as evident by increase in PARP cleavage (**Figure 11B**). Similarly, cells expressing shRNA that targets PRCAT47 formed smaller tumors in mice compared to cells expressing non-targeting shRNA (**Figure 11C**). Taken together, we show that PRCAT47 is important for survival of tumor cells.

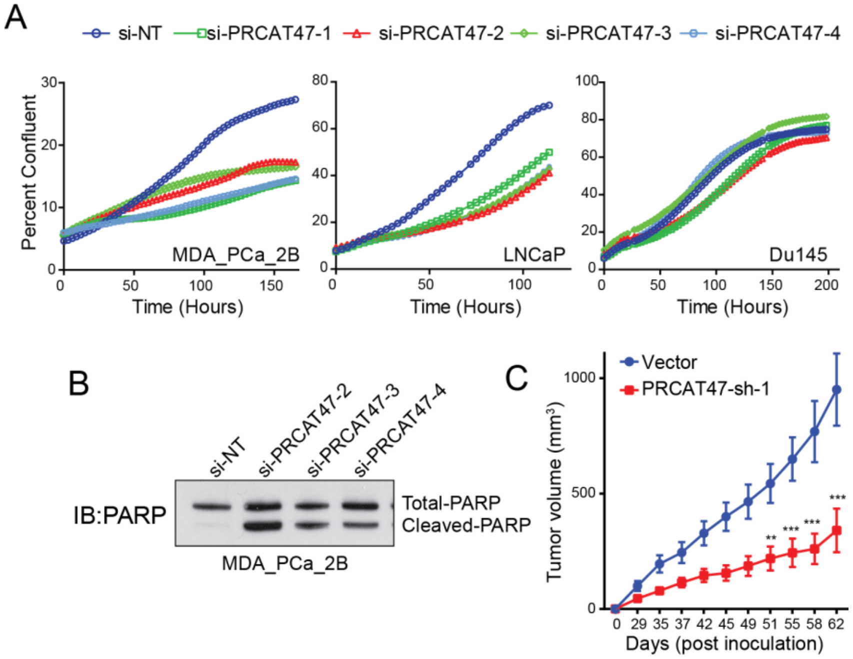


Figure 11: A-C. Effect of PRCAT47 KD on cell proliferation (A), apoptosis (B) and *in vivo* tumor growth (C).

Specific Aim 2: To elucidate the function of lncRNA in metastatic castrate resistant prostate cancer (Months 6-24)

The following aim was designed to initiate studies that can inform on the mechanism of action of lncRNA function. We proposed to perform gene expression analysis by microarrays (Task-1), Correlation analysis (Task-4) as well as experiment to identify the protein interactome of lncRNA (Task 2-4). We hypothesize that these studies together will be instrumental in understanding the mechanism of lncRNA.

To begin to understand the mechanism of action of PRCAT47, we performed guilt by association studies. Briefly, we identified the genes that positively and negatively correlated with PRCAT47 using the TCGA prostate cancer RNA-seq data. We next performed GSEA and oncomine analysis on the correlated genes to identify potential concepts that could guide us towards mechanism. As shown in **Figure 12**, below correlated genes were significantly enriched in prostate cancer concepts suggesting role of PRCAT47 in cancer progression.

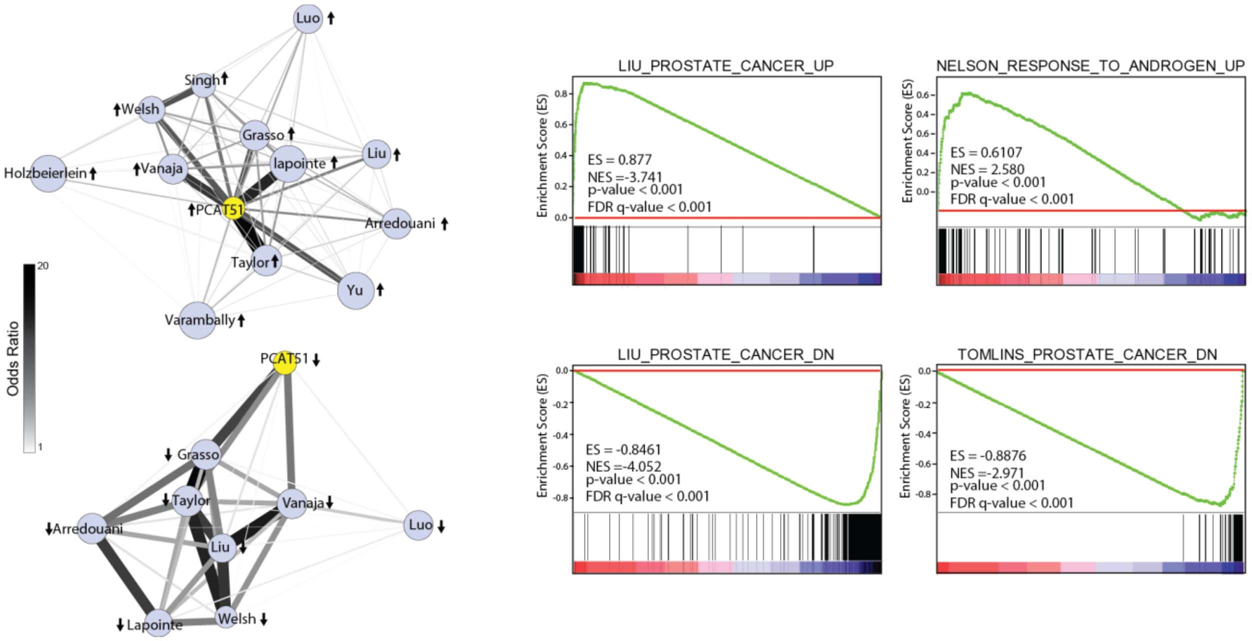
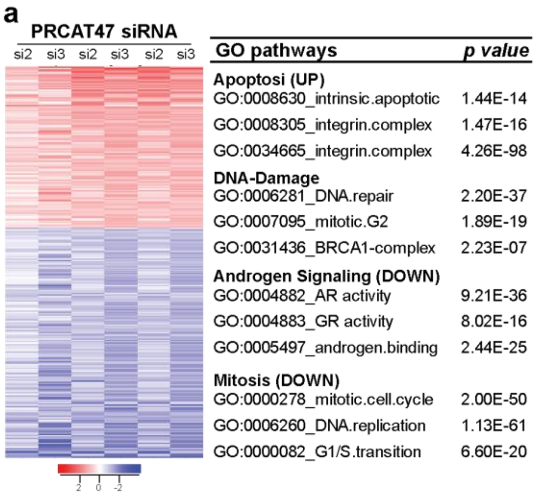


Figure 12: Cytoscape network representation of Oncomine concepts analysis of genes correlated with PRCAT47 in localized prostate cancers profiled by RNA-Seq. Gene set enrichment analysis (GSEA) showing significant enrichment of genes involved in prostate cancer

To further identify the function of PRCAT47 in prostate cancer, we performed gene expression profiling of MDA PCa-2b cells treated with siRNAs targeting PRCAT47. Upon GO analysis of the gene expression data, we found that PRCAT47 KD regulated genes involved in cell proliferation and apoptosis (**Figure 13A**).

Interestingly, in gene expression analysis of PRCAT47 KD cells, we also noticed a significant decrease in AR target gene expression (**Figure 13A**), suggesting a positive feedback loop between PRCAT47 and AR. To confirm this observation, we generated an AR target gene signature from MDA PCa-2b cells stimulated with DHT and performed GSEA analysis using this gene set. As expected, knockdown of PRCAT47 led to decreased expression of genes positively regulated by AR and increased expression of gene negatively regulated by AR (**Figure 14A).** This was further confirmed by qPCR and western blot analysis of AR target genes upon PRCAT47 knockdown (**Figure 14B-D**).



GO pathways	*p value*
Apoptosi (UP)	
GO:0008630_intrinsic.apoptotic	1.44E-14
GO:0008305_integrin.complex	1.47E-16
GO:0034665_integrin.complex	4.26E-98
DNA-Damage	
GO:0006281_DNA.repair	2.20E-37
GO:0007095_mitotic.G2	1.89E-19
GO:0031436_BRCA1-complex	2.23E-07
Androgen Signaling (DOWN)	
GO:0004882_AR activity	9.21E-36
GO:0004883_GR activity	8.02E-16
GO:0005497_androgen.binding	2.44E-25
Mitosis (DOWN)	
GO:0000278_mitotic.cell.cycle	2.00E-50
GO:0006260_DNA.replication	1.13E-61
GO:0000082_G1/S.transition	6.60E-20

Figure 13: Effect of PRCAT47 KD on global gene expression. GO pathway analysis was performed on gene expression data obtained after PRCAT47 knockdown.

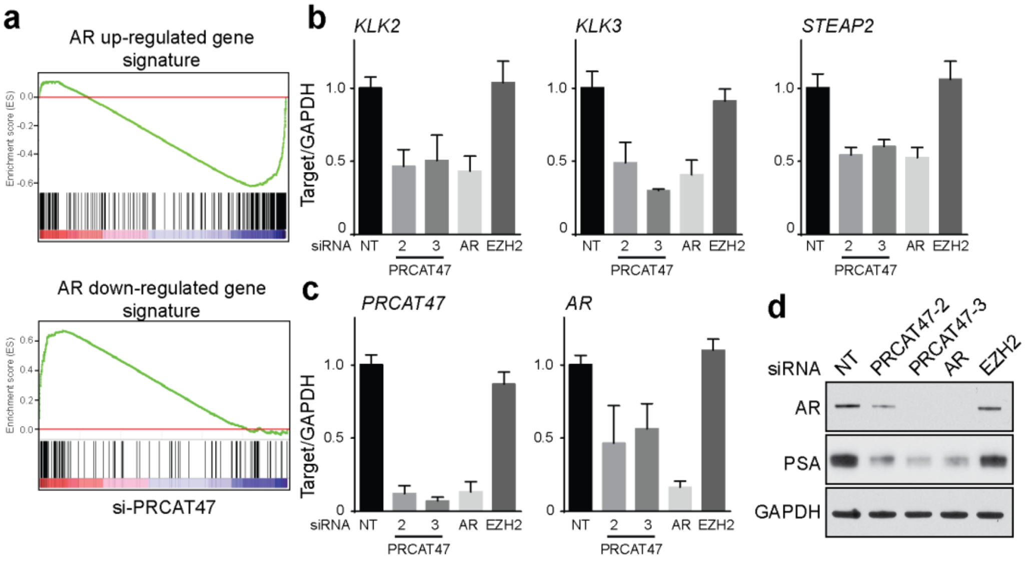


Figure 14: a. GSEA against AR target gene set using gene expression data from **Figure 13**. Effect of PRCAT47 knockdown on canonical AR target genes. **c**. Effect of PRCAT47 KD on AR mRNA and vice-versa. **d.** Effect of PRCAT47 KD on protein levels of AR and PSA.

To begin characterization of the potential mechanism through which *PRCAT47* is functioning in cells, we investigated the cellular localization of *PRCAT47*, observing it to be present in both the cytoplasm and nucleus via qRT-PCR following cellular fractionation (**Figure-15).** With this pan-cellular expression pattern, we hypothesized that *PRCAT47* may be executing its function through a protein partner and set out to identify potential protein interactors via RNA-pulldown followed by mass-spectrometry.

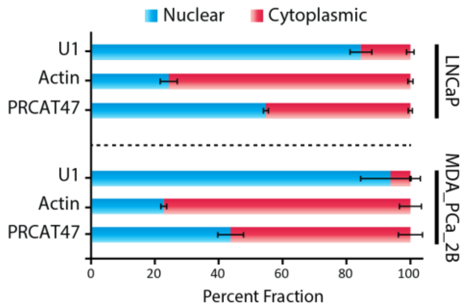
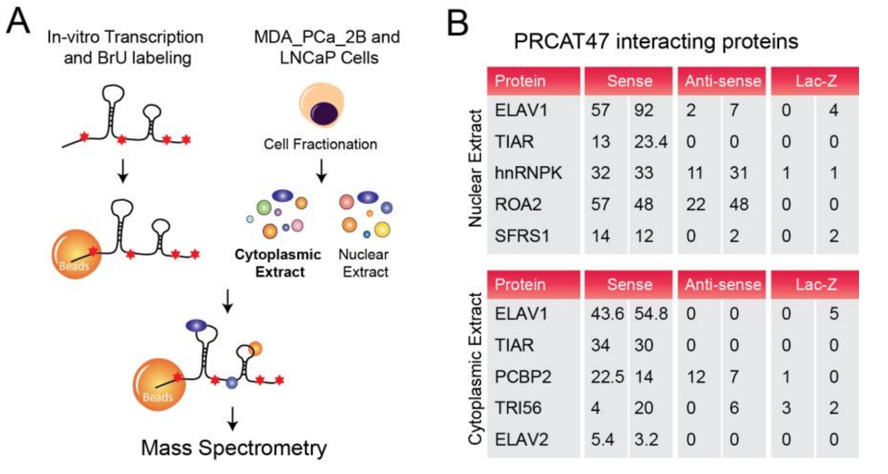


Figure-15. LNCaP and MDA_PCa_2B cells were fractionated and the expression of PRCAT47was analyzed by qRT-PCR

We initially proposed to perform ChIRP mass spec, where an endogenous RNA is pulled down using biotinylated probes. We designed 10 tiling probes spanning PRCAT47 gene and tested their ability to pulldown endogenous PRCAT47 from cell lysates by qPCR on bound RNA. However, the amount of protein that was pulldown with the RNA was not enough for Mass-Spec analysis. Furthermore, the fixing reagent required for ChIRP was not compatible for Mass-Spec analysis. We decided to approach this problem using an alternative protocol. In order to identify RNA interacting proteins we performed *in-vitro* RNA pull-downs following mass-spectrometry.



PRCAT47 interacting proteins

	Protein	Sense		Anti-sense		Lac-Z	
Nuclear Extract	ELAV1	57	92	2	7	0	4
	TIAR	13	23.4	0	0	0	0
	hnRNPK	32	33	11	31	1	1
	ROA2	57	48	22	48	0	0
	SFRS1	14	12	0	2	0	2

	Protein	Sense		Anti-sense		Lac-Z	
Cytoplasmic Extract	ELAV1	43.6	54.8	0	0	0	5
	TIAR	34	30	0	0	0	0
	PCBP2	22.5	14	12	7	1	0
	TRI56	4	20	0	6	3	2
	ELAV2	5.4	3.2	0	0	0	0

Figure 16: A. Schematic representation of the BrU pulldown approach. **B.** List of protein in nucleus and cytoplasm that were found to be associated with PRCAT47.

Briefly, full length PRCAT47 RNA was synthesized *in-vitro* and labeled with BrU. Labeled RNA was then incubated with cell lysates, pulled down using anti-BrU antibodies and bound proteins were analyzed by mass spectrometery. Antisense of PRCAT47 and lac-z RNA were used as negative controls. We found many proteins that interact with PRCAT47. HuR and TIAR proteins that bind to AU-rich elements on RNA, were found to be the top ranking proteins that interact with PRCAT47 (**Figure 16B**). We further confirmed these interactions by performing RNA immunoprecipitation (RIP) using HuR and TIAR antibodies followed by quantification of PRCAT47 by qPCR (**Figure 17A,B**).

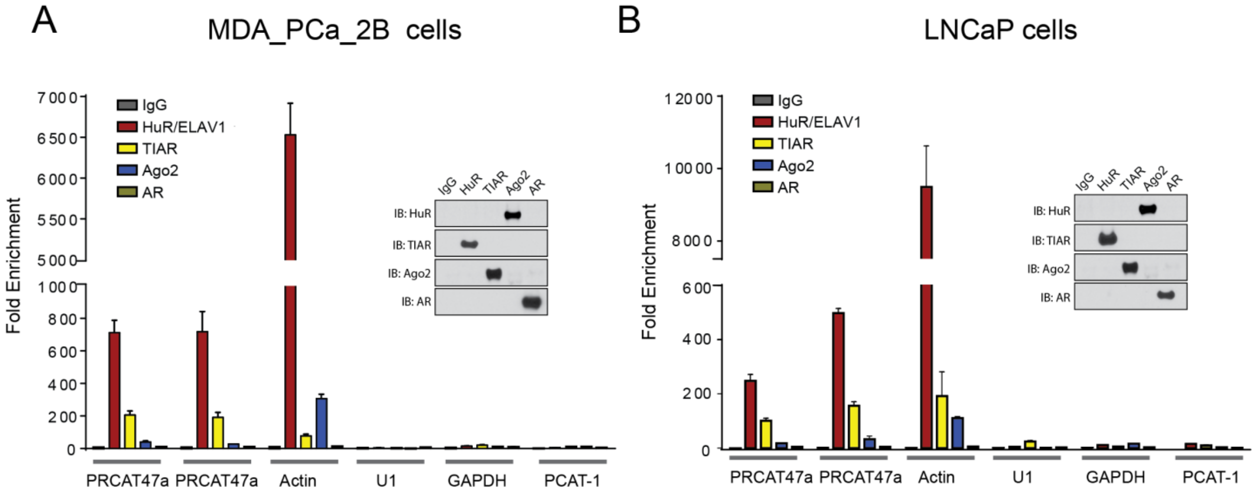


Figure 17: RNA immunprecipitation (RIP) was performed to validate the binding of PRCAT47 with candidate protein in MDA_PCa_2B (A) and LNCaP (B) cells.

A detailed mechanistic insight would be required to understand how PRCAT47 interacts with HuR and TIAR to control the levels of AR in cells. A possible hypothesis would be that PRCAT47 bridges HuR and TIAR to AR mRNA. This interaction would stabilize AR mRNA and would lead to higher AR protein in cells.

PCAT29: To understand the mechanism of PCAT29 function, we designed two independent shRNAs to knock down the expression of PCAT29 in cells VCaP and LNCaP cells (cell that express PCAT29 at high levels) and performed gene expression microarray. We found GO concepts enriched for cell cycle, proliferation, and migration related genes, suggesting a role of PCAT29 in proliferation and migration (**Figure 18**).

Next, we performed correlation analysis, similar to PRCAT47, and defined a signature of genes positively and negatively correlated with PCAT29 expression from prostate cancer samples (7). Interestingly, apart from prostate cancer vs normal concepts, we were able to also identify concepts of genes that were involved in EMT (Figure 19). This was consistent with our observation that PCAT29 acts like a tumor suppressor. However, a

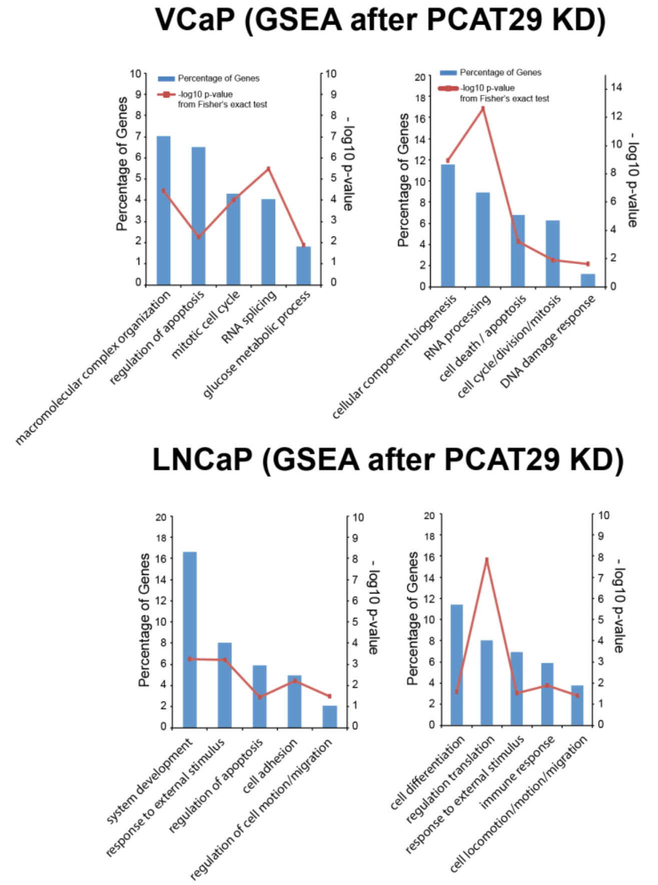


Figure 18: GSEA analysis of micorarray data obtained after PCAT29 knockdown in VCaP and LNCaP cells. Bars represent the Geo categories enriched.

detailed molecular insight into the mechanism would be required to further understand the role of PCAT29 in prostate cancer.

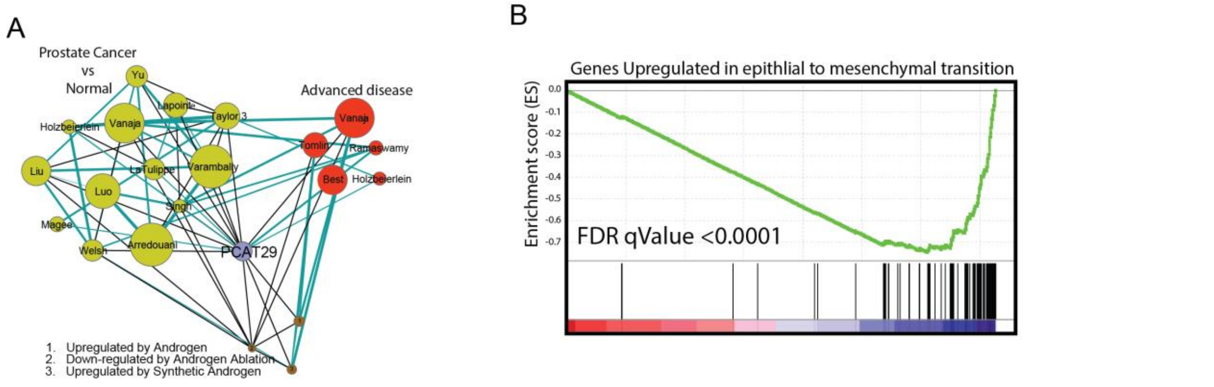


Figure 19: A. Cytoscape network representation of Oncomine concepts analysis of genes correlated with PCAT29 expression levels in localized prostate cancers profiled by RNA-Seq. **B**. Gene set enrichment analysis (GSEA) showing significant enrichment of genes involved in epithelial to mesenchymal transition.

Training: I did training with bioinformatician in lab and learned how to analyze microarray data. I learned to use LIMMA package in R for data analysis. Other software's such as MeV4 and Treeview were also learned for heat map generation. Further, I am taking courses on "R" with the University of Michigan biostatistics department.

Specific Aim 3: To identify lncRNAs that serve as potential biomarkers of disease progression. (Months 12-24).

This aim was designed to explore the role of lncRNAs as potential biomarkers of disease initiation (diagnostic) or progression (progression). For a gene, including lncRNA, to be a clinically important biomarkers (prognostic or diagnostic), it is important that **1)** It is differentially expressed in cancer compared to normal tissue, **2)**. It is tissue- and cancer-specific and **3)**. Must validate in multiple cohorts. To address this, we have assembled multiple prostate RNA-seq cohorts. Furthermore, we have downloaded and analyzed the TCGA RNA-seq data (18 different cancer types, approximately 6800 samples) to assess the tissue specificity of particular lncRNA. Upon examining the expression of PCAT29 and

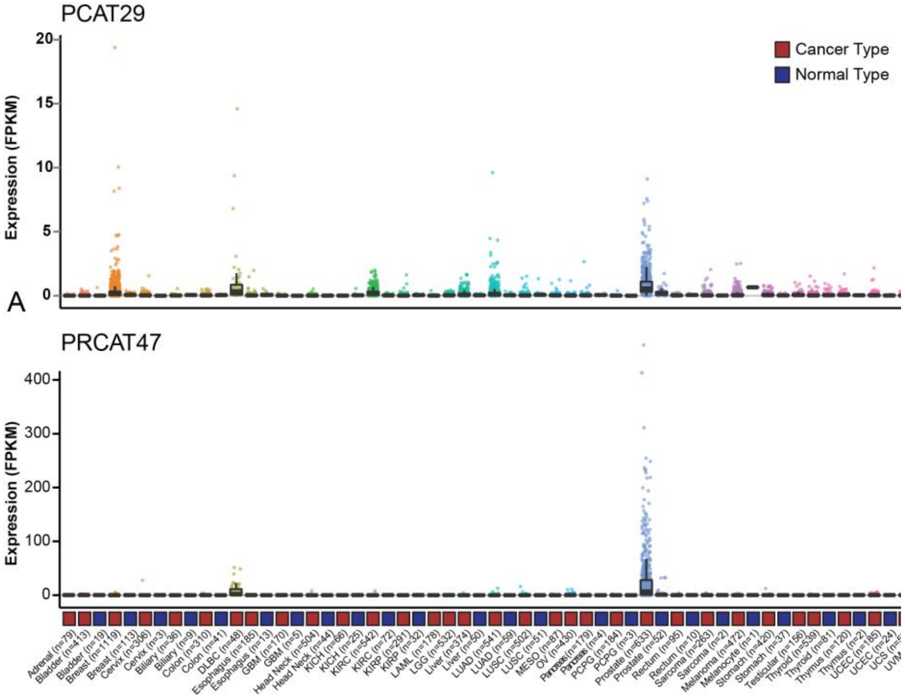


Figure 20 : Expression of PCAT29 and PRCAT47 in various tissues as determined by RNA-Seq

PRCAT47 in this compendium of samples, we observed that, although PCAT29 was differentially expressed in prostate cancer compared to benign prostate, its expression was not restricted to prostate cancer. Several other tumor types also expressed PCAT29 at high levels (**Figure 20**).

In contrast, expression of PRCAT47 was mostly restricted to prostate cancer, making it a better biomarker candidate than PCAT29. Similar to PRCA47, SChLAP1 has also been shown to be a prostate cancer- and tissue-specific marker. SChLAP-1 expression predicts for poor survival of prostate cancer patients [5, 7]. Based on this we further examined the biomarker potential of PRCAT47.

Task 1: Screening of lncRNA in prostate cancer cohorts (Months 12-24):
Progress (100% completed)
We screened multiple prostate cancer cohorts for the expression of PRCAT47. As shown in Figure 21, in all the cohorts examined, PRCAT47 expression levels were found to be upregulated in both localized and metastatic prostate cancer compared to benign tissues.

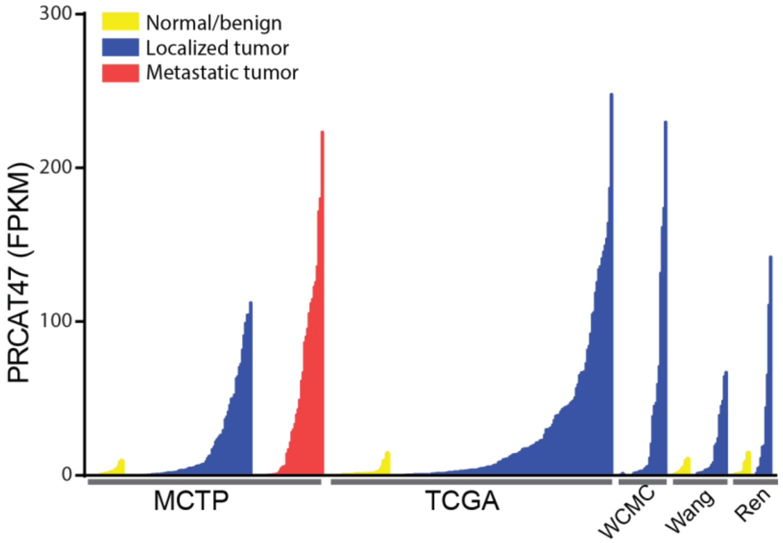


Fig 21: Expression of PRCAT47 in MCTP (Michigan center for Translational Pathology, TCGA (The Cancer genome Atlas), WCMC (Weil college of Medicine), and other published cohorts [8, 9].

Task 2 and 3: Screening of lncRNAs in FFPE and Fresh frozen tissue cohort (Months 12-24).
Progress (100% completed)
LncRNA detection in cancer tissue sections by RNA *in-situ* hybridization (RNA-ISH) technology has similar clinical utility as immunohistochemical evaluation of protein biomarkers. Our lab, using SChLAP-1 as an example, has shown the clinical utility of this approach [10, 11]. This approach can be applied to formalin fixed paraffin-embedded (FFPE) tissues which provides an opportunity to validate our findings in larger cohorts with associated clinical data in the future. We initially proposed to perform qPCR on FFPE and fresh frozen cohorts, however, because of our recent success in detecting lncRNA such as SChLAP-1 in FFPE cohorts of prostate cancer, we decided to screen PRCAT47 using RNA-ISH approach. We focused on PRCAT47 as it was highly expressed in prostate cancer tissues.

In order to screen the FFPE cohorts, we first made RNA-ISH probes that could specifically recognize PRCAT47 in FFPE fixed tissues. Development of RNA-ISH probes as well as staining of FFPE tissue microarrays (TMA) cohorts was done by ACD technologies. We first assessed the expression of PRCAT47 in FFPE sections made from xenografts of MDA-PCa-2B cells, a cell line that expresses PRCAT47 at very high levels. As expected, ISH probes were able to detect PRCAT47 in these xenograft sections (**Figure 22A**). No staining was seen in other normal tissue, suggesting that the probes are specific. We next performed RNA *in-situ* hybridization of PRCAT47 in prostate TMA comprised of benign tissues, localized and metastatic prostate cancer samples. As expected, expression of PRCAT47 was found to be high in localized and metastatic tissues compared to the benign samples **(Figure 22B, C**).

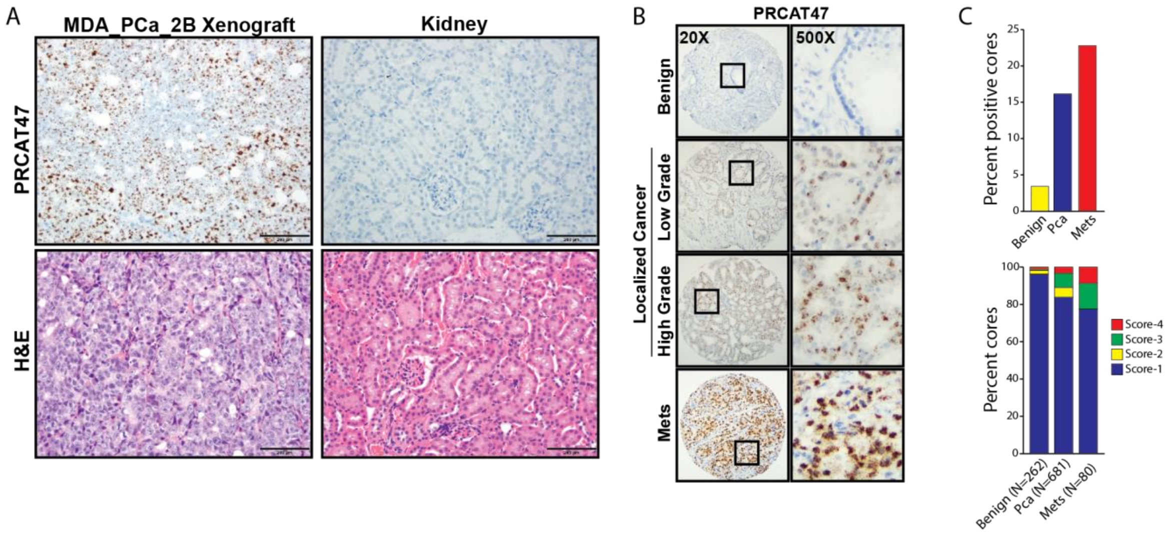


Fig 22: A. Expression of PRCAT47 by RNA ISH in MDA-PCa-2b xenografts sections as well as mouse kidney. **B**. Expression of PRCAT47 in benign human prostate and cancer sections of varying Gleason Grade. **C.** Quantification of PRCAT47 staining in Tissue microarrays comprised of prostate cancer and benign samples.

Task 4: Microarray analysis of lncRNA in cohorts with clinical outcome (Months 18-24):

We assessed the expression of PRCAT47 in Mayo clinic cohort of prostate cancer samples [5]. The expression was determined using Affymetrix Human Exon ST 1.0 microarray. As shown in figure 23, PRCAT47 expression was able to predict for freedom from biochemical recurrence. As expected, high PRCAT47 lead to higher rates of biochemical recurrence.

,

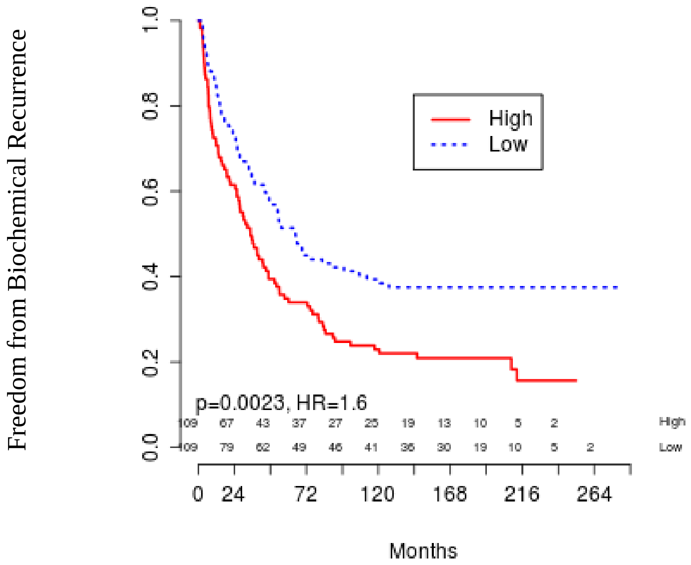


Fig. 23: Kaplan Mayer analysis of PRCAT47 expression to predict for biochemical recurrence.

4.KEY RESEARCH ACCOMPLISHMENTS:

In course of this proposal several key research milestones were accomplished. Some of the major accomplishments are highlighted below.

a. We identified and characterized PCAT29 as first AR suppressed tumor suppressor lncRNA in prostate cancer.

b. Using RNA-Seq approach, we defined a compendium of lncRNAs that are regulated by AR as well as differentially regulated in prostate cancer. This will provide us opportunity to identify and characterize novel lncRNAs that may play role in prostate cancer pathogenesis and may have important biomarker capabilities.

c. We identified a novel lncRNA (PRCAT47) that is critical for prostate cancer progression by regulation AR signaling. This provides us unique opportunity to use lncRNA as therapeutic targets.

5.CONCLUSION:

In the past few years, there has been a drastic increase in the discovery and characterization of long non-coding RNAs associated with a variety of disorders including cancers. Our group recently identified thousands of novel lncRNAs across multiple cancer types using RNA-Seq data from more than 7000 cancer samples coupled with a newly developed analysis platform. However, our understanding of the mechanism of action of lncRNAs is still incomplete with only a hand full of lncRNAs characterized mechanistically. This post-doctoral fellowship from the DOD helped in achieving these goals at many fronts.

Going forward, I would like to continue exploring the role of lncRNAs in prostate tumor progression as well as their potential as biomarker and therapeutic candidates. Two specific areas of research that I am particularly interested are

1) **To explore the therapeutic potential of lncRNAs**. With advances in anti-sense oligo's technology, it is now possible to target RNA molecules for therapeutic purposes. In future, I would like to explore possibility of targeting lncRNAs in prostate cancer. Based on data generated from this proposal, PRCAT47 seems to be a good target for therapeutics as it regulates prostate cell proliferation.

2) **To identify novel biomarkers for disease progression**: LncRNAs are particularly interesting biomarker candidates because they show both disease and tissue specific expression. In future, I would like to identify new lncRNA candidates that can act as biomarkers of disease progression.

I have written a DOD idea development award to pursue these research goals. This award has been funded and will help me to explore lncRNAs further in the context of prostate cancer.

6. PUBLICATIONS, ABSTRACTS, AND PRESENTATIONS:

Peer-Reviewed Scientific Journals

a. Prensner, J.R., M.K. Iyer, A. Sahu, I.A. Asangani, Q. Cao, L. Patel, I.A. Vergara, E. Davicioni, N. Erho, M. Ghadessi, R.B. Jenkins, T.J. Triche, R. Malik, R. Bedenis, N. McGregor, T. Ma, W. Chen, S. Han, X. Jing, X. Cao, X. Wang, B. Chandler, W. Yan, J. Siddiqui, L.P. Kunju, S.M. Dhanasekaran, K.J. Pienta, F.Y. Feng, and A.M. Chinnaiyan. The long noncoding RNA SChLAP1 promotes aggressive prostate cancer and antagonizes the SWI/SNF complex. Nat Genet. 2013 Nov;45(11):1392-8. PMCID: PMC3812362

b. Prensner, J.R.*, A. Sahu*, M.K. Iyer*, Malik R.*, B. Chandler, I.A. Asangani, A. Poliakov, I.A. Vergara, M. Alshalalfa, R.B. Jenkins, E. Davicioni, F.Y. Feng, and A.M. Chinnaiyan. The lncRNAs PCGEM1 and PRNCR1 are not implicated in castration resistant prostate cancer. Oncotarget. 2014 Mar 30;5(6):1434-8 (*Co-first Author) PMCID: PMC4039221

c. Malik R., Patel L., Prensner, J.R, Shi Y., Iyer M., Subramaniyan S., Alexander Carley, Yashar S. Niknafs, Anirban Sahu, Sumin Han, Teng Ma, Meilan Liu, Irfan Asangani, Xiaojun Jing, Xuhong Cao, Mohan Dhaneshekaran, Dan Robinson, Felix Y. Feng, Arul M. Chinnaiyan. The lncRNA PCAT29 Inhibits Oncogenic Phenotypes in Prostate Cancer. Mol. Can. Res. 2014 Aug;12(8):1081-7. PMCID: PMC4135019

d. Iyer MK*, Niknafs YS*, Malik R, Singhal U, Sahu A, Hosono Y, Barrette TR, Prensner JR, Evans JR, Zhao S, Poliakov A, Cao X, Dhanasekaran SM, Wu YM, Robinson DR, Beer DG, Feng FY, Iyer HK, Chinnaiyan AM. The landscape of long noncoding RNAs in the human transcriptome. Nat Genet. 2015 Mar;47(3):199-208. (*Co-first Author) PMCID: 25599403

Lay Press

a. Matthew Iyer, **Rohit Malik**, Anirban Sahu, Javed Siddiqui, Arul Chinnaiyan. *qPCR Validation of Novel Prostate-specific Long Non-coding RNAs.* Application Note; Wafergene Biosciences; 2013. http://www.wafergen.com/wp-content/uploads/2013/01/UM_lncRNA_TNf.pdf

Invited Articles: Nothing to report

Abstracts: Nothing to report

Meeting abstracts are reported below

b. List presentations made during the last year (international, national, local societies, military meetings, etc.). Use an asterisk (*) if presentation produced a manuscript.

a. **Rohit Malik**, Yajia Zhang, Yashar Niknafs, Matthew K Iyer, Marcin Cieslik, Yasuyuki Hosono, Shruthi Subramaniam, Sahr Yazdani, Anirban Sahu, Xuhong Cao, Dan Robinson, Felix Feng and Arul M Chinnaiyan. In: Proceedings of the Keystone Symposium; "Long Noncoding RNA: From Evolution to function". March 15-20th, Keystone, CO.

b. **Rohit Malik**, Matthew K Iyer, Shruthi Subramaniam, Yasuyuki Hosono, Anirban Sahu, Xia Jiang, Yang Shi, Vishal Kothari, Xuhong Cao, Dan Robinson, Saravana M. Dhanasekaran, Felix Y Feng and Arul M Chinnaiyan**. Investigating the Biological Function of an Androgen-Receptor Upregulated**

Long noncoding RNA (ARUL-1) in Prostate Cancer. Keystone Symposia Conference. "Long Noncoding RNAs: Marching toward Mechanism" 2014

c. **Rohit Malik,** Matthew K. Iyer, John R. Prensner, Lalit Patel, Sumin Han, Wei Chen, Felix Feng, Arul M. Chinnaiyan. **Identification and characterization of a novel androgen-regulated long non-coding RNA in prostate cancer**. In: Proceedings of the 104th Annual Meeting of the American Association for Cancer Research; 2013 Apr 6-10; Washington, DC. Philadelphia (PA): AACR; ***Cancer Res*** 2013;73 (8 Suppl):Abstract nr 1120. doi:10.1158/1538-7445.AM2013-1120

d. **Rohit Malik**, Matthew K. Iyer, John R. Prensner, Lalit Patel, Sumin Han, Wei Chen, Felix Feng, Arul M. Chinnaiyan. "Sixth Annual Prostate Cancer Program Retreat". March 18-20, 2013

7. INVENTIONS, PATENTS AND LICENSES:

Nothing to report

8. REPORTABLE OUTCOMES:

Nothing to report

9. OTHER ACHIEVEMENTS:

Post-Doctoral funding from the Department of Defense has been very instrument for my career. Based on funding and publication, I was promoted to Research Investigator position in the Department of Pathology at the University of Michigan.

Based on the results generated from this DOD postdoctoral award, I recently applied for a DOD Idea Development award that has been recommended for funding. The title of this proposal is "**Discovery and characterization of PRCAT47: A novel prostate lineage and cancer specific long non-coding RNA**".

10. REFERENCES:

1. Malik, R., et al., *The lncRNA PCAT29 inhibits oncogenic phenotypes in prostate cancer.* Mol Cancer Res, 2014. **12**(8): p. 1081-7.
2. Robinson, D., et al., *Integrative clinical genomics of advanced prostate cancer.* Cell, 2015. **161**(5): p. 1215-28.
3. Grasso, C.S., et al., *The mutational landscape of lethal castration-resistant prostate cancer.* Nature, 2012. **487**(7406): p. 239-43.
4. Iyer, M.K., et al., *The landscape of long noncoding RNAs in the human transcriptome.* Nat Genet, 2015. **47**(3): p. 199-208.
5. Prensner, J.R., et al., *The long noncoding RNA SChLAP1 promotes aggressive prostate cancer and antagonizes the SWI/SNF complex.* Nat Genet, 2013. **45**(11): p. 1392-8.
6. Prensner, J.R., et al., *Transcriptome sequencing across a prostate cancer cohort identifies PCAT-1, an unannotated lincRNA implicated in disease progression.* Nat Biotechnol, 2011. **29**(8): p. 742-9.

7. Prensner, J.R., et al., *RNA biomarkers associated with metastatic progression in prostate cancer: a multi-institutional high-throughput analysis of SChLAP1.* Lancet Oncol, 2014. **15**(13): p. 1469-80.
8. Ren, S., et al., *RNA-seq analysis of prostate cancer in the Chinese population identifies recurrent gene fusions, cancer-associated long noncoding RNAs and aberrant alternative splicings.* Cell Res, 2012. **22**(5): p. 806-21.
9. Zhai, W., et al., *Transcriptome profiling of prostate tumor and matched normal samples by RNA-Seq.* Eur Rev Med Pharmacol Sci, 2014. **18**(9): p. 1354-60.
10. Mehra, R., et al., *A novel RNA in situ hybridization assay for the long noncoding RNA SChLAP1 predicts poor clinical outcome after radical prostatectomy in clinically localized prostate cancer.* Neoplasia, 2014. **16**(12): p. 1121-7.
11. Mehra, R., et al., *Overexpression of the Long Non-coding RNA SChLAP1 Independently Predicts Lethal Prostate Cancer.* Eur Urol, 2015.

11. APPENDICES: N/A

TRAINING OR FELLOWSHIP AWARDS: N/A

Opportunities for training and professional development

Over the Department of Defense Postdoctoral Fellowship funding period, I engaged in numerous training and professional development activities. This included presentations (both oral and poster) at several scientific conferences, writing grants and manuscripts, one-on-one meeting with my mentor to discuss scientific progress and future plans as well as mentoring and teaching students.

Presentations at Scientific Conference: The data that was collected as part of this project was presented at several scientific conferences. Listed below are the conferences that I attended where I presented my work:

1. AACR Special Conference on Noncoding RNAs and Cancer: Mechanisms to Medicines, December 4-7, 2015
2. Keystone Symposium; “Long Noncoding RNA: From Evolution to function”. March 15-20th, 2015, Keystone, CO.
3. Keystone Symposia; “Long Noncoding RNAs: Marching toward Mechanism” 2014
4. 104th Annual Meeting of the American Association for Cancer Research; 2013 Apr 6-10;
5. Sixth Annual Prostate Cancer Program Retreat; March 18-20, 2013

Invited Talks at conferences: My abstracts were selected for oral presentations at several conferences. I was also invited to give oral presentations at two upcoming meetings. By giving these talks on my lncRNA research, I am able to refine my presentation skills as well as learn to present the data in a clear and succinct fashion. Some of the oral presentations that I have given or will be giving are listed below:

1. American Society for Investigative Pathology. Workshop: Long Non-Coding RNA. April 4-6, 2016
2. Ninth Annual Prostate Cancer Program Retreat. March 13-15, 2016
3. University of Michigan “RNA Super group” meeting
4. Keystone Symposia; “Long Noncoding RNAs: Marching toward Mechanism” 2014

5. 104th Annual Meeting of the American Association for Cancer Research; 2013 Apr 6-10;
6. Sixth Annual Prostate Cancer Program Retreat; March 18-20, 2013

Manuscript writing: The work from this project resulted in several manuscripts that I developed. A manuscript describing the studies of lncRNA PCAT29 was published in 2014 (Mol Cancer Res, 2014. **12**(8): p. 1081-7). Another manuscript describing the results of the study on PRCAT47 is currently under preparation.

Grant Writing

With input from the mentor (Dr. Chinnaiyan) I also developed and submitted various grants. The results obtained from this proposal were utilized as preliminary data to write a DOD IDEA development award. This award that is likely to be funded will play a major role in my career development.

Mentoring

Constant feedback and advice from the mentor was provided both at several levels.

> **Weekly lab meetings:** All research findings were presented during weekly lab meetings in front of the lab group. A group meeting focused specifically on long non coding RNA research is held every week. These meeting provide opportunity to present my data to the group and get insights from the mentor.
>
> **One-on-one meeting with mentor**: In addition to presenting the data at the group meetings, data is discussed with mentor (Dr. Arul Chinnaiyan) in biweekly individual meetings and *via* monthly progress reports. Constant inputs and advice were provided. These meetings are of tremendous help as they help guide the direction of the project.

Mentoring Experience

I continued to mentor several individuals. As part of the UROP (University Research opportunity program) at the University of Michigan, I mentored several undergraduate students. I also help train research technicians as well as graduate students.

Published OnlineFirst July 16, 2014; DOI: 10.1158/1541-7786.MCR-14-0257

Molecular Cancer Research

The lncRNA *PCAT29* Inhibits Oncogenic Phenotypes in Prostate Cancer

Rohit Malik, Lalit Patel, John R. Prensner, et al.

Mol Cancer Res 2014;12:1081-1087. Published OnlineFirst July 16, 2014.

Published OnlineFirst July 16, 2014; DOI: 10.1158/1541-7786.MCR-14-0257

Molecular Cancer Research

MCR RapidIMPACT

The lncRNA *PCAT29* Inhibits Oncogenic Phenotypes in Prostate Cancer

Rohit Malik[1,2], Lalit Patel[1,2], John R. Prensner[1,2], Yang Shi[1,3], Matthew K. Iyer[1,2], Shruthi Subramaniyan[1], Alexander Carley[1], Yashar S. Niknafs[1,2], Anirban Sahu[1,2], Sumin Han[1,4], Teng Ma[1,4], Meilan Liu[4], Irfan A. Asangani[1,2], Xiaojun Jing[1,2], Xuhong Cao[1,2], Saravana M. Dhanasekaran[1,2], Dan R. Robinson[1,2], Felix Y. Feng[1,4,5], and Arul M. Chinnaiyan[1,2,5,6]

Abstract

Long noncoding RNAs (lncRNA) have recently been associated with the development and progression of a variety of human cancers. However, to date, the interplay between known oncogenic or tumor-suppressive events and lncRNAs has not been well described. Here, the novel lncRNA, prostate cancer–associated transcript 29 (*PCAT29*), is characterized along with its relationship to the androgen receptor. *PCAT29* is suppressed by DHT and upregulated upon castration therapy in a prostate cancer xenograft model. *PCAT29* knockdown significantly increased proliferation and migration of prostate cancer cells, whereas *PCAT29* overexpression conferred the opposite effect and suppressed growth and metastases of prostate tumors in chick chorioallantoic membrane assays. Finally, in prostate cancer patient specimens, low *PCAT29* expression correlated with poor prognostic outcomes. Taken together, these data expose *PCAT29* as an androgen-regulated tumor suppressor in prostate cancer.

Implications: This study identifies *PCAT29* as the first androgen receptor–repressed lncRNA that functions as a tumor suppressor and that its loss may identify a subset of patients at higher risk for disease recurrence.

Visual Overview: http://mcr.aacrjournals.org/content/early/2014/07/31/1541-7786.MCR-14-0257/F1.large.jpg.

Mol Cancer Res; 12(8); 1081–7. ©2014 AACR.

Introduction

Recently, data from the ENCODE project have revealed that the majority of the transcriptome is composed of noncoding RNAs (1). While the classification of these noncoding RNAs is still in development, long noncoding RNAs (lncRNA) are of particular interest, given the similar features they share with protein-coding genes as well as recent evidence of their roles in cancer biology (2, 3). LncRNAs are polyadenylated RNA species that are more than 200 bp in length, transcribed by RNA polymerase II, and associated with common epigenetic signatures such as of histone 3 lysine 4 trimethylation (H3K4me3) at the transcriptional start site (TSS) and histone 3 lysine 36 trimethylation (H3K36me3) in the gene body (4). Several lncRNAs have been shown to play a role in biologic processes such as X-chromosomal inactivation, pluripotency (5), and gene regulation (6). Recently, several lncRNAs have been implicated in cancer initiation and progression (3, 7). Apart from their role in tumor initiation and progression, lncRNAs have been shown to be promising biomarkers. In prostate cancer, PCA3 is a well-studied prostate cancer biomarker that is now available for clinical use as a urine biomarker assay for diagnosis of prostate cancer (8, 9).

Despite their involvement in various cellular processes, the majority of lncRNAs are uncharacterized, and their role in cancer initiation and progression remains unclear. Using transcriptome sequencing, our group recently identified more than 100 lncRNAs, named prostate cancer–associated transcripts (*PCATs*), which are differentially expressed or have outlier profiles in prostate cancer versus normal tissue (3). Here we find that one of these novel lncRNAs, *PCAT29*, exhibits cancer-suppressive phenotypes, including inhibition of cell proliferation, migration, tumor growth, and metastases. In accordance with this, *PCAT29* is repressed by androgen signaling, and low *PCAT29* expression associates with worse clinical outcomes.

[1]Michigan Center for Translational Pathology, University of Michigan, Ann Arbor, Michigan. [2]Department of Pathology, University of Michigan, Ann Arbor, Michigan. [3]Department of Biostatistics, University of Michigan, Ann Arbor, Michigan. [4]Department of Radiation Oncology, University of Michigan, Ann Arbor, Michigan. [5]Comprehensive Cancer Center, University of Michigan, Ann Arbor, Michigan. [6]Howard Hughes Medical Institute, University of Michigan, Ann Arbor, Michigan.

Note: Supplementary data for this article are available at Molecular Cancer Research Online (http://mcr.aacrjournals.org/).

F.Y. Feng and A.M. Chinnaiyan share senior authorship of this article.

Corresponding Authors: Arul M. Chinnaiyan, Comprehensive Cancer Center, University of Michigan Medical School, 1400 E. Medical Center Dr. 5316 CCGC 5940, Ann Arbor, MI 48109. Phone: 734-615-4062; Fax: 734-615-4055; E-mail: arul@med.umich.edu; and Felix Y. Feng, Department of Radiation Oncology, University of Michigan Medical Center, 1500 East Medical Center Drive, UHB2C490-SPC5010, Ann Arbor, MI 48109. Phone: 734-936-4302; Fax: 734-763-7371; E-mail: ffeng@med.umich.edu

doi: 10.1158/1541-7786.MCR-14-0257

Published OnlineFirst July 16, 2014; DOI: 10.1158/1541-7786.MCR-14-0257

Materials and Methods

Cell lines and reagents

Prostate cancer cells were cultured as follows: VCaP cells in DMEM with GlutaMAX (Invitrogen) and LNCaP and DU145 cells in RPMI-1640 (Invitrogen) in a 5% CO2 cell culture incubator. All the media were supplemented with 10% FBS (Invitrogen) and 1% penicillin–streptomycin (Invitrogen). All cell lines were purchased from ATCC and were authenticated.

For stable knockdown of *PCAT29*, LNCaP and VCaP cells were transfected with lentiviral constructs encoding 2 different *PCAT29* shRNAs or nontargeting shRNAs in the presence of polybrene (8 μg/mL; Supplementary Table S1A). After 48 hours, transduced cells were grown in culture media containing 3 to 5 μg/mL puromycin. For PCAT29 overexpression, 2 isoforms of *PCAT29* were generated by subcloning the PCR product into the *CPO1* sites of the pCDH-CMV vector (System Biosciences). Five hundred base pairs of the genomic region was attached at the 5′ end of each isoform. Lentiviral particles were made and DU145 cells were transduced as described above.

Gene expression by quantitative PCR

Total RNA was isolated using TRIzol (Invitrogen) and an RNeasy kit (Qiagen) according to manufacturers' instruction. Total RNA was reverse transcribed into cDNA using SuperScript III and random primers (Invitrogen). Quantitative PCR (qPCR) was performed using SYBR Green Master Mix (Applied Biosystems) on an Applied Biosystems 7900HT Real-Time System. The relative quantity of the target gene was computed for each sample using the $\Delta\Delta C_t$ method by comparing mean C_t of the gene to the mean C_t of the housekeeping gene *GAPDH*. All the primers were obtained from Integrated DNA Technologies (IDT). Sequences of all the primers used are listed in Supplementary Table S1B.

Rapid amplification of cDNA ends)

5′ and 3′ RACE was performed using the GeneRacer RLM-RACE kit (Invitrogen) following manufacturer's instruction. RACE PCR products were separated on a 1% agarose gel. Individual bands were gel purified, cloned in pcr4-TOPO vector, and sequenced using M13 primers.

Expression of *PCAT29* after castration in prostate tumor xenograft model

Five-week-old male nude athymic BALB/c *nu*/*nu* mice (Charles River Laboratory) were used for xenograft studies. LNCaP cells were resuspended in 100 μL of PBS with 20% Matrigel (BD Biosciences) and implanted subcutaneously into the left flank regions of the mice. Mice were castrated and euthanized 5 days after castration. RNA was extracted from the xenografts and expression of *PCAT29* and *FKBP5* was measured. All experimental procedures involving mice were approved by the University Committee on Use and Care of Animals at the University of Michigan (Ann Arbor, MI) and conform to their relevant regulatory standards.

Chromatin immunoprecipitation

Chromatin immunoprecipitation (ChIP) was performed with polyclonal androgen receptor antibody (Millipore PG21) using HiCell ChIP kit (Diagenode) following manufacturer's instruction. Briefly, cells were treated with 10 μmol/L MDV3100 or 10 μmol/L bicalutamide 16 hours before the treatment with 10 nmol/L DHT for 12 hours. Approximately 1 million cells were cross-linked per antibody with 1% formaldehyde. Chromatin was sonicated to an average length of 500 bp and centrifuged to remove debris. Magnetic protein-G beads were coated with 6 μg of antibody and incubated with chromatin overnight at 4°C. Protein–chromatin–antibody complexes were washed thrice and cross-linking was reversed. ChIP products were cleaned using IPure kit (Diagenode). Eluted DNA was quantified by RT-PCR using primers described in Supplementary Table S1B.

Cell proliferation and migration assay

LNCaP and DU145 cells stably expressing *PCAT29* shRNA-1 and 2 or *PCAT29* isoform 1 and 2 were seeded in 24-well plates. Cells were trypsinized and counted by using Coulter Counter (Beckman Coulter) at the indicated time points in triplicate. For migration assays, approximately 1×10^5 cells were seeded in the upper chamber of a Boyden chamber. About 500 μL of complete medium (10% FBS) was added to the lower chamber as a chemoattractant. Forty-eight hours after seeding, cells on the upper surface were removed using a cotton swab. Inserts were fixed with 3.7% formaldehyde and migrated cells on the lower surface of the membrane were stained with crystal violet. The inserts were treated with 10% acetic acid, and absorbance was measured at 560 nm.

Gene expression microarray

Expression profiling of VCaP and LNCaP cells after *PCAT29* knockdown was performed using the Agilent Whole Human Genome Oligo Microarray as described (7). GEO accession number: GSE58397.

Chicken chorioallantoic membrane assay

22RV1 cells were transduced with empty vector (pcDH) or *PCAT29*-isoform-1. A total of 10^6 cells were inoculated on the chicken chorioallantoic membrane (CAM) assay as described previously (10). For tumor growth and metastasis, the eggs were incubated for 18 days in total, after which the extra-embryonic tumor were exercised and weighed, and the embryonic livers were harvested and analyzed for the presence of tumor cells by quantitative human Alu-specific PCR. Quantification of human cells in the extracted DNA was performed as described (11). Fluorogenic TaqMan qPCR probes were applied as above and DNA copy numbers were quantified.

Kaplan–Meier analysis of *PCAT29*

For outcomes analysis, *PCAT29* expression was determined on a cohort of 51 radical prostatectomy specimens from patients with prostate cancer at the University of

Published OnlineFirst July 16, 2014; DOI: 10.1158/1541-7786.MCR-14-0257

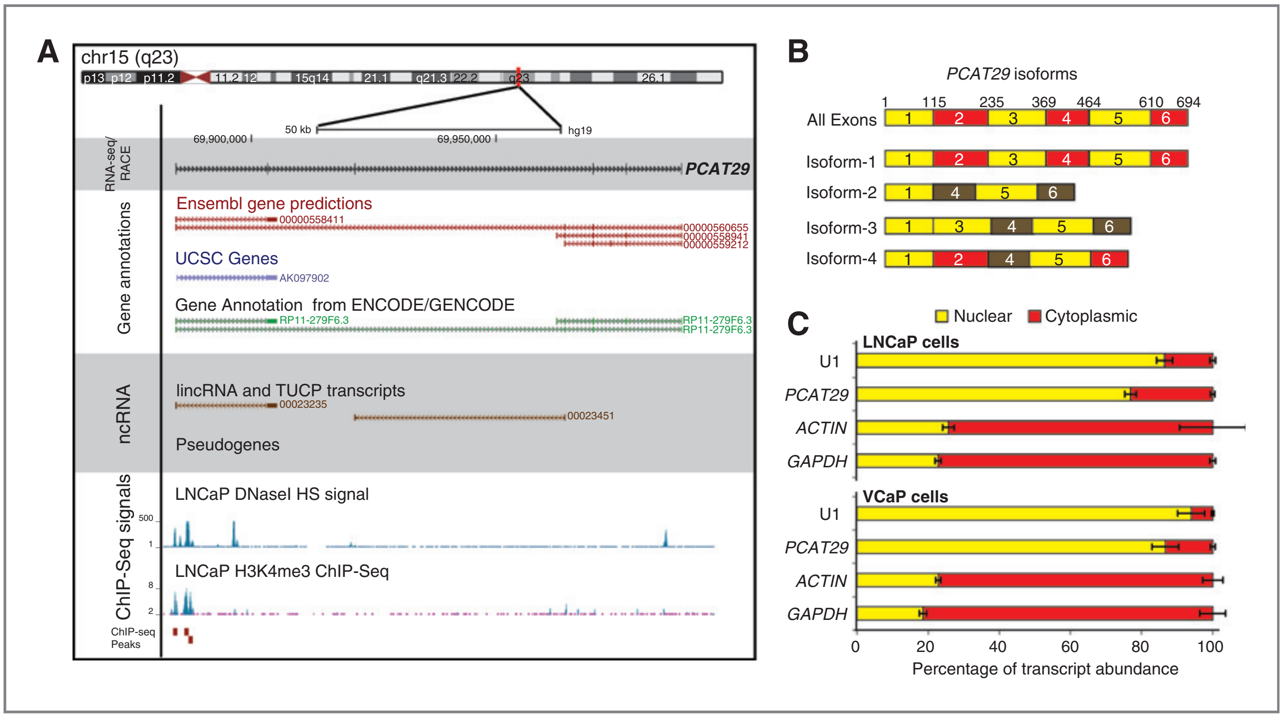


Figure 1. Characterization of *PCAT29*. A, genome browser representation of *PCAT29*. Current gene annotations from Ensembl, ENCODE, UCSC genes and lncRNA databases are shown. ChIP-Seq data for H3K4me3 and DNaseI HS signal in LNCaP cells obtained from ENCODE. B, schematic representation of PCAT29 isoforms as determined by RACE analysis. C, nuclear and cytoplasmic distribution of various noncoding and protein-coding transcripts in LNCaP and VCaP cells. Error bar, ±SEM.

Michigan with long-term biochemical recurrence outcomes. Biochemical recurrence was defined by an increase of PSA of 0.2 ng/mL over the PSA nadir following prostatectomy. *PCAT29* expression was determined by a SYBR-Green qPCR assay using the average of *GAPDH* + *HMBS* for data normalization using the $\Delta\Delta C_t$ method. Expression data were transformed using a *z*-score and patients were defined as high (top 33% of patients) or low (bottom 66% of patients) for *PCAT29* expression. Kaplan–Meier curves for biochemical recurrence-free survival were generated for *PCAT29*-high and *PCAT29*-low patients using the GraphPad Prism program. Statistical significance was determined with a log-rank test.

Results

PCAT29 is a novel long nuclear noncoding RNA

Using RNA-Seq data from prostate cancer tissues, we previously identified 121 lncRNAs, named PCATs, which demonstrate differential expression or outlier profiles in prostate cancer compared with normal tissue (3). Here we characterize and functionally investigate one of the top outlier lncRNAs, *PCAT29* (Ensembl ID ENSG00000259641). Using the predicted transcript structures, we designed exon spanning primers and performed rapid amplification of cDNA ends (RACE) to determine the full exon structure. As shown in a genome browser view, *PCAT29* is a 694-bp polyadenylated transcript present on chr15(q23), and the *PCAT29* gene spans over a 10-kb stretch (Fig. 1A; Supplementary Fig. S1A). *PCAT29* is composed of 6 exons that are alternatively spliced to produce multiple isoforms (Fig. 1B). To further characterize *PCAT29*, we interrogated recently published ENCODE data for H3K4 trimethylation (H3K4me3) and DNaseI hypersensitive sites (DNaseH), marks that predicts for open chromatin state and are commonly found near or at the TSSs, generated in the prostate cancer cell line LNCaP (4). We found several DNaseH and H3K4 trimethylation peaks at the TSS of *PCAT29*, suggesting that *PCAT29* is an actively transcribed gene (Fig. 1A).

To confirm that PCAT29 is indeed a noncoding RNA, we assessed the protein-coding potential of *PCAT29* using the coding potential calculator (CPC) algorithm, which discriminates coding genes (positive score) from noncoding transcripts (negative score; ref. 12). *PCAT29* had a CPC score of −0.8921, whereas protein-coding genes such as *TP53* and *β-actin* scored +8.25 and +3.70, respectively (Supplementary Fig. S1B). Consistent with this finding, we found that in both LNCaP and VCaP cells, expression of *PCAT29* was limited to nucleus, whereas other protein-coding mRNAs, such as *GAPDH* and *β-actin*, were expressed in cytoplasm (Fig. 1C). We then verified the expression of *PCAT29* in various prostate cancer cell lines (LNCaP, VCaP, 22RV1, DU145, PC3) and immortalized or primary prostate epithelial cells (RPWE and PrEC). *PCAT29* expression was highest in androgen receptor–dependent cell lines such as LNCaP, VCaP, and 22RV1 (Supplementary Fig. S1C).

Published OnlineFirst July 16, 2014; DOI: 10.1158/1541-7786.MCR-14-0257

Next, we assessed the expression of *PCAT29* in various tissues using transcriptome sequencing data. *PCAT29* expression, although not limited to prostate, was enriched in prostate samples compared with other tissues (Supplementary Fig. S1D).

Androgen receptor binds to the *PCAT29* promoter and regulates *PCAT29* expression

We next examined the effect of androgen receptor signaling on *PCAT29* in LNCaP cells stimulated with 10 nmol/L DHT. As shown in Fig. 2A, *PCAT29* expression was suppressed upon stimulation with DHT in a time-dependent fashion both in LNCaP and VCaP cells. In contrast, expression of canonical androgen receptor target genes, such as *FKBP5* and *KLK3*, was increased upon stimulation (Supplementary Fig. S2A). To examine whether the suppression of *PCAT29* was androgen receptor–specific, LNCaP cells were pretreated with the androgen receptor antagonists MDV3100 or bicalutamide before treatment with DHT. As expected, DHT stimulation suppressed the expression of *PCAT29*, and pretreatment with MDV3100 or bicalutamide rescued this suppression. Similarly, expression of *PCAT29* in LNCaP cells grown in charcoal-stripped media as well as in an androgen receptor–independent variant of LNCaP cells (C42) was higher than in cells grown in serum-containing media and LNCaP cells, respectively (Supplementary Fig. S2B and S2C). We next investigated whether androgen receptor suppresses the expression of *PCAT29 in vivo*. LNCaP xenografts were established in mice followed by physical castration to ablate androgen receptor signaling. As expected, 5 days of castration led to significant increase in the expression of *PCAT29* in tumors (Fig. 2C). In contrast, expression of *FKBP5* was reduced in tumors from castrated mice. Taken together, our results suggest that stimulation

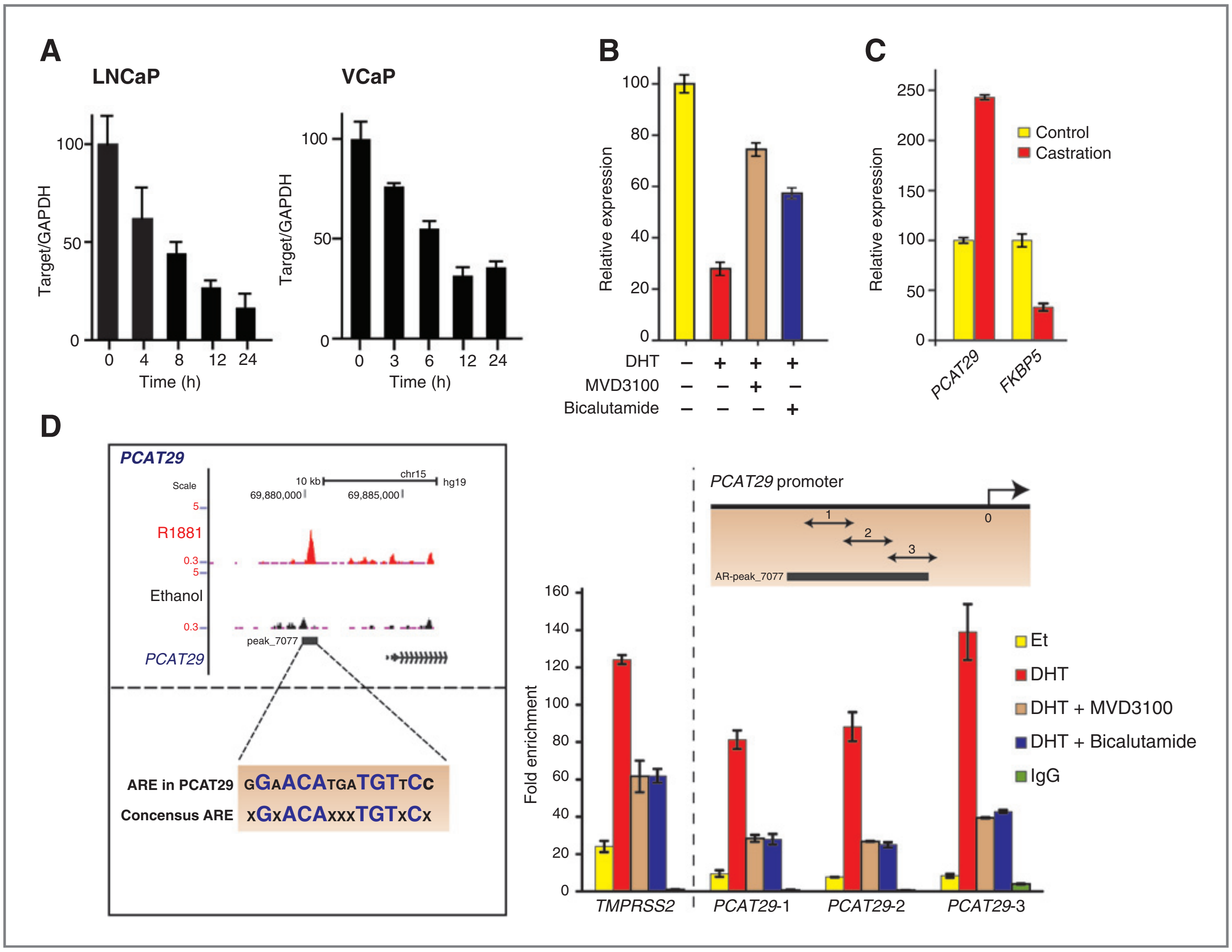


Figure 2. Androgen receptor binds to the promoter of *PCAT29* and suppresses its expression. A, expression of *PCAT29* in LNCaP and VCaP cells treated with 10 nmol/L DHT for indicated time points. B, expression of *PCAT29* in LNCaP cells treated with 10 nmol/L DHT in the presence or absence of 10 μmol/L MDV3100 or bicalutamide. C, expression of *PCAT29* and *FKBP5* in LNCaP xenografts obtained from control mice and mice that were physically castrated for 5 days. D, genome browser representation of androgen receptor (AR) binding on the promoter of *PCAT29* before and after stimulation with 1 nmol/L R1881. Consensus androgen-responsive elements (ARE) and ARE present in the *PCAT29* promoter are shown. Inset, ChIP-PCR to confirm AR occupancy on *TMPRSS2* and *PCAT29* gene promoter. The *y*-axis represents AR ChIP enrichment in VCaP cells treated with 10 nmol/L DHT normalized to ethanol (Ethl)-treated cells. Bars, SEM.

Published OnlineFirst July 16, 2014; DOI: 10.1158/1541-7786.MCR-14-0257

of androgen receptor leads to suppression of *PCAT29* expression.

To further study the association of *PCAT29* expression with androgen signaling, we interrogated published ChIP-Seq data (13) and found androgen receptor–binding sites in the promoter region of *PCAT29* (Fig. 2D). These peaks were similar to those observed in other known androgen receptor–regulated genes (Supplementary Fig. S2D). Upon closer inspection, we found a canonical androgen receptor–binding site near the *PCAT29* TSS in a putative enhancer region bounder by androgen receptor (Fig. 2D). We confirmed our ChIP-Seq data by performing ChIP for androgen receptor followed by PCR for the *PCAT29* promoter. As shown in Fig. 2D, stimulation of VCaP cells with DHT led to an increase in association of androgen receptor with the *PCAT29* promoter. This association was reduced in cells pretreated with bicalutamide and MDV3100. Taken together, our data suggest that androgen receptor can directly bind to the promoter of *PCAT29* and leads to the suppression of gene expression.

PCAT29 regulates oncogenic phenotypes *in vitro* and *in vivo*

The androgen receptor drives oncogenesis in treatment-naïve prostate cancer as well as disease progression in castration-resistant prostate cancers. Because androgen receptor binds to the *PCAT29* promoter and regulates gene expression, we investigated the functional role of *PCAT29*. Two independent shRNAs were designed to knockdown the expression of *PCAT29* in cells (Supplementary Fig. S3A and S3B). VCaP and LNCaP cells were transfected with *PCAT29* shRNAs following analysis using gene expression microarray. We found GO concepts enriched for cell cycle, proliferation, and migration-related genes, suggesting a role of *PCAT29* in proliferation and migration (Supplementary Fig. S3D–S3G). Next, we defined a signature of genes positively and negatively correlated with *PCAT29* expression from prostate cancer samples as described before (7). We checked the overlap of these genes with the top 1500 differentially expressed genes in *PCAT29* knockdown samples of VCAP and LNCAP cells. As expected, the positively correlated genes show a significant overlap with genes downregulated with knockdown of *PCAT29* and the negatively correlated genes show a significant overlap with genes upregulated by knockdown of *PCAT29* in both VCAP and LNCAP ($P < 0.001$ for all pairwise comparisons of overlapping genes, Supplementary Fig. S4A–S4D). For overlapping genes, we did see enrichment in pathways such as cell cycle, apoptosis, and cell growth (Supplementary Fig. S4A–S4D). Taken together, this analysis suggested a role of *PCAT29* in cell proliferation and migration.

To experimentally validate this observation, cell proliferation was assessed in LNCaP cells transfected with control versus *PCAT29* shRNAs. To our surprise, knockdown of *PCAT29* in LNCaP cells led to an increase in cell proliferation and migration (Fig. 3A). To further validate this observation, we stably overexpressed the 2 most prevalent isoforms of *PCAT29* in DU145 prostate cancer cells using a lentiviral vector (Supplementary Fig. S3C). Consistent with the previous knockdown studies, overexpression of these 2 isoforms of *PCAT29* in DU145 led to suppression of cell proliferation and migration (Fig. 3B). We next assessed whether similar effects of *PCAT29* could be achieved *in vivo*. 22RV1 prostate cancer cells overexpressing *PCAT29* (isoform-1) were implanted on the CAM of a chicken egg. Compared with control cells, overexpression of *PCAT29* significantly decreased the growth of tumor on the CAM as well as decreased liver metastases (Fig. 3C).

Finally, we measured the expression of *PCAT29* in an independent cohort of 51 radical prostatectomy specimens from patients with prostate cancer with localized disease and clinical follow-up. As shown in Kaplan–Meier analysis (Fig. 3D), patients with lower *PCAT29* expression had significantly higher rates of biochemical recurrence, consistent with our *in vitro* and *in vivo* findings.

Discussion

In this study, we characterize the novel lncRNA *PCAT29*. Our findings demonstrate that PCAT29 is directly regulated by the androgen receptor, which binds to the promoter of *PCAT29* and suppresses its transcription. *In vitro* studies show that *PCAT29* negatively regulates prostate cancer proliferation and migration, and CAM assays demonstrate that *PCAT29* inhibits tumor growth and metastases. Low expression of *PCAT29* is associated with higher rates of biochemical recurrence, suggesting that *PCAT29* represses oncogenic phenotypes via a tumor-suppressive role.

While previous studies have nominated and characterized lncRNAs that are dysregulated in cancer (3, 14), the majority of characterized lncRNAs, to date, has been associated with oncogenic roles instead of tumor suppressor functions. In fact, there have been only a handful of lncRNAs identified to date that function in repression of cancer phenotypes, and, to our knowledge, none of these are targets downregulated by known oncogenes (14). A recent study identifies a protein-coding gene, CCN3/NOV, as a tumor suppressor that is repressed by androgen receptor (15). Thus, our study represents the first identification of an androgen receptor–repressed lncRNA functioning as a tumor suppressor. While further studies will be required to identify the mechanism of *PCAT29* and other tumor suppressor lncRNAs, it is clear that these lncRNAs represent an intriguing area for exploration in cancer biology.

In the context of prostate cancer, androgen-regulated lncRNAs are of fundamental importance, given that all stages of prostate cancer are exquisitely dependent on androgen receptor signaling for growth and survival. Because the majority of clinically relevant prostate cancer therapies target the androgen receptor, our studies would suggest that inhibition of androgen signaling will result in reactivation of *PCAT29*, providing another mechanism underlying the effectiveness of androgen deprivation therapy.

Clinically, there is a clear need for identification of prognostic biomarkers in prostate cancer to help guide decisions on treatment intensification. The association of high *PCAT29* expression with good clinical prognosis and

Published OnlineFirst July 16, 2014; DOI: 10.1158/1541-7786.MCR-14-0257

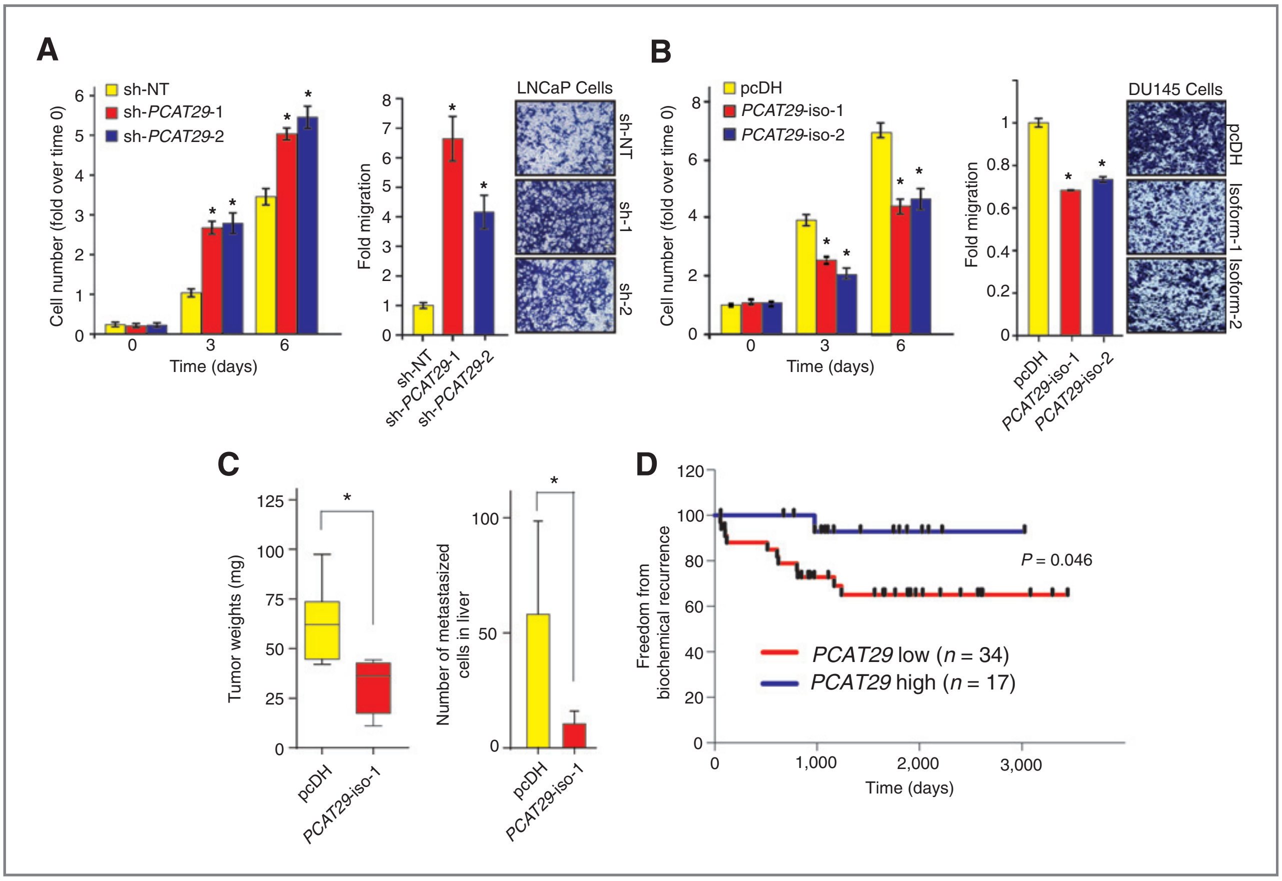


Figure 3. *PCAT29* suppresses oncogenic phenotypes. A and B, proliferation and migration of LNCaP cells stably expressing *PCAT29* shRNA (B) and DU145 cells expressing *PCAT29* expression constructs (C). Representative micrographs of crystal violet–stained migrated cells are shown as insets. C, quantification of tumor weight and metastasis to liver for 22Rv1 cells expressing *PCAT29*-isoform 1 or empty vector (pcDH) in the CAM assay. Data, mean ± SEM. *, $P < 0.05$ by the Student *t* test. D, Kaplan–Meier analyses of prostate cancer outcomes. *PCAT29* expression was measured by qPCR and 51 patients were stratified according to their *PCAT29* expression. Patient outcomes were analyzed for freedom from biochemical recurrence.

preclinical suppression of cell proliferation and tumor metastases suggests that decreased expression or loss of *PCAT29* may identify subsets of patients who may require further intensification of therapy. As our clinical cohort was composed of hormone-sensitive disease from patients with prostatectomy, further studies need to be performed to determine whether *PCAT29* can also serve as a prognostic biomarker in the context of more advanced, castration-resistant disease. Regardless, this study highlights the importance of lncRNAs in prostate cancer biology and prognosis and suggests the need for further research in this relatively unexplored area.

Disclosure of Potential Conflicts of Interest

J.R. Prensner has ownership interest as a co-inventor on prostate cancer ncRNA patent licensed to GenomeDx Biosciences Inc., including PCATs in prostate cancer. M. Iyer has ownership interest in a patent (ncRNA and uses thereof). A.M. Chinnaiyan is a consultant/advisory board member for Wafergen Inc. No potential conflicts of interest were disclosed by the other authors.

Authors' Contributions

Conception and design: R. Malik, D.R. Robinson, F.Y. Feng, A.M. Chinnaiyan
Development of methodology: R. Malik, L. Patel, J.R. Prensner, M.K. Iyer, I.A. Asangani, X. Jing, X. Cao, A.M. Chinnaiyan
Acquisition of data (provided animals, acquired and managed patients, provided facilities, etc.): R. Malik, L. Patel, J.R. Prensner, S. Subramaniyan, A. Carley, A. Sahu, S. Han, M. Liu, I.A. Asangani, X. Jing, X. Cao, S.M. Dhanasekaran, D.R. Robinson, F.Y. Feng
Analysis and interpretation of data (e.g., statistical analysis, biostatistics, computational analysis): R. Malik, L. Patel, J.R. Prensner, Y. Shi, M.K. Iyer, A. Carley, Y.S. Niknafs, A. Sahu, I.A. Asangani, S.M. Dhanasekaran, D.R. Robinson, F.Y. Feng
Writing, review, and/or revision of the manuscript: R. Malik, J.R. Prensner, D.R. Robinson, F.Y. Feng, A.M. Chinnaiyan
Administrative, technical, or material support (i.e., reporting or organizing data, constructing databases): M.K. Iyer, T. Ma, X. Jing, X. Cao, D.R. Robinson, F.Y. Feng
Study supervision: R. Malik, D.R. Robinson, F.Y. Feng, A.M. Chinnaiyan

Acknowledgments

The authors thank Xia Jiang and Vishal Kothari for helpful discussions and technical assistance, as well as Karen Giles for her review of the manuscript and submission of documents. They also thank the University of Michigan Viral vector core for generating the lentiviral constructs.

Grant Support

This work was supported in part by the NIH Prostate Specialized Program of Research Excellence Grant P50CA69568, the Early Detection Research Network grant UO1 CA111275, the U.S. NIH R01CA132874-01A1, and the Department of Defense grant PC100171 (A.M. Chinnaiyan). A.M. Chinnaiyan is supported by the Doris Duke Charitable Foundation Clinical Scientist Award, the Prostate Cancer Foundation, and the Howard Hughes Medical Institute. A.M. Chinnaiyan is an American Cancer Society Research Professor and a Taubman Scholar of the

University of Michigan. F.Y. Feng was supported by the Prostate Cancer Foundation and the Department of Defense grant PC094231. R. Malik and J.R. Prensner were supported by a Prostate Cancer Foundation Young Investigator Award. R. Malik was supported by the Department of Defense Postdoctoral Fellowship W81XWH-13-1-0284. A. Sahu was supported by the NIH Predoctoral Fellowship 1F30CA180376-01. M. Iyer was supported by the Department of Defense Predoctoral Fellowship BC100238.

Received May 8, 2014; revised June 17, 2014; accepted June 26, 2014; published OnlineFirst July 16, 2014.

References

1. Djebali S, Davis CA, Merkel A, Dobin A, Lassmann T, Mortazavi A, et al. Landscape of transcription in human cells. Nature 2012;489:101–8.
2. Du Z, Fei T, Verhaak RG, Su Z, Zhang Y, Brown M, et al. Integrative genomic analyses reveal clinically relevant long noncoding RNAs in human cancer. Nat Struct Mol Biol 2013;20:908–13.
3. Prensner JR, Chinnaiyan AM. The emergence of lncRNAs in cancer biology. Cancer Discov 2011;1:391–407.
4. Guttman M, Amit I, Garber M, French C, Lin MF, Feldser D, et al. Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. Nature 2009;458:223–7.
5. Loewer S, Cabili MN, Guttman M, Loh YH, Thomas K, Park IH, et al. Large intergenic non-coding RNA-RoR modulates reprogramming of human induced pluripotent stem cells. Nat Genet 2010;42:1113–7.
6. Tsai MC, Manor O, Wan Y, Mosammaparast N, Wang JK, Lan F, et al. Long noncoding RNA as modular scaffold of histone modification complexes. Science 2010;329:689–93.
7. Prensner JR, Iyer MK, Sahu A, Asangani IA, Cao Q, Patel L, et al. The long noncoding RNA SChLAP1 promotes aggressive prostate cancer and antagonizes the SWI/SNF complex. Nat Genet 2013;45:1392–8.
8. Tomlins SA, Aubin SM, Siddiqui J, Lonigro RJ, Sefton-Miller L, Miick S, et al. Urine TMPRSS2:ERG fusion transcript stratifies prostate cancer risk in men with elevated serum PSA. Sci Transl Med 2011;3:94ra72.
9. Lee GL, Dobi A, Srivastava S. Prostate cancer: diagnostic performance of the PCA3 urine test. Nat Rev Urol 2011;8:123–4.
10. Asangani IA, Ateeq B, Cao Q, Dodson L, Pandhi M, Kunju LP, et al. Characterization of the EZH2-MMSET histone methyltransferase regulatory axis in cancer. Mol Cell 2013;49:80–93.
11. van der Horst EH, Leupold JH, Schubbert R, Ullrich A, Allgayer H. TaqMan-based quantification of invasive cells in the chick embryo metastasis assay. Biotechniques 2004;37:940–2, 4, 6.
12. Kong L, Zhang Y, Ye ZQ, Liu XQ, Zhao SQ, Wei L, et al. CPC: assess the protein-coding potential of transcripts using sequence features and support vector machine. Nucleic Acids Res 2007;35:W345–9.
13. Yu J, Mani RS, Cao Q, Brenner CJ, Cao X, Wang X, et al. An integrated network of androgen receptor, polycomb, and TMPRSS2-ERG gene fusions in prostate cancer progression. Cancer Cell 2010;17:443–54.
14. Nie L, Wu HJ, Hsu JM, Chang SS, Labaff AM, Li CW, et al. Long non-coding RNAs: versatile master regulators of gene expression and crucial players in cancer. Am J Transl Res 2012;4:127–50.
15. Wu L, Runkle C, Jin HJ, Yu J, Li J, Yang X, et al. CCN3/NOV gene expression in human prostate cancer is directly suppressed by the androgen receptor. Oncogene 2014;33:504–13

nature genetics

ANALYSIS

The landscape of long noncoding RNAs in the human transcriptome

Matthew K Iyer[1,2,11], Yashar S Niknafs[1,3,11], Rohit Malik[1,4], Udit Singhal[1,5], Anirban Sahu[1,4], Yasuyuki Hosono[1], Terrence R Barrette[1], John R Prensner[1], Joseph R Evans[1,6], Shuang Zhao[1,6], Anton Poliakov[1], Xuhong Cao[1,5], Saravana M Dhanasekaran[1,4], Yi-Mi Wu[1], Dan R Robinson[1], David G Beer[6,7], Felix Y Feng[1,6,8], Hariharan K Iyer[9] & Arul M Chinnaiyan[1,2,4,5,8,10]

Long noncoding RNAs (lncRNAs) are emerging as important regulators of tissue physiology and disease processes including cancer. To delineate genome-wide lncRNA expression, we curated 7,256 RNA sequencing (RNA-seq) libraries from tumors, normal tissues and cell lines comprising over 43 Tb of sequence from 25 independent studies. We applied *ab initio* assembly methodology to this data set, yielding a consensus human transcriptome of 91,013 expressed genes. Over 68% (58,648) of genes were classified as lncRNAs, of which 79% were previously unannotated. About 1% (597) of the lncRNAs harbored ultraconserved elements, and 7% (3,900) overlapped disease-associated SNPs. To prioritize lineage-specific, disease-associated lncRNA expression, we employed non-parametric differential expression testing and nominated 7,942 lineage- or cancer-associated lncRNA genes. The lncRNA landscape characterized here may shed light on normal biology and cancer pathogenesis and may be valuable for future biomarker development.

Cancers are a leading cause of morbidity and mortality worldwide, with over 14 million new cases and 8 million deaths in 2012 (ref. 1). To improve understanding of cancer pathogenesis, ongoing large-scale efforts led by The Cancer Genome Atlas (TCGA) are using high-throughput molecular profiling strategies to characterize genetic, epigenetic and transcriptional changes[2,3]. However, efforts to interpret these data have mainly focused on protein-coding genes, despite definitive evidence that transcription of the noncoding genome produces functional RNAs[4]. In particular, lncRNAs have been implicated in biological, developmental and pathological processes and act through mechanisms such as chromatin reprogramming, *cis* regulation at enhancers and post-transcriptional regulation of mRNA processing[5,6].

The emergence of high-throughput RNA-seq technology provides a revolutionary means for the systematic discovery of transcriptional units. Indeed, RNA-seq has led to a deeper appreciation of the intricate nature of transcription by identifying a milieu of lncRNAs, both located in intergenic 'gene deserts' and overlapping protein-coding loci[4]. The aligned sequence data generated by RNA-seq experiments can be used to predict full-length transcripts *in silico* with *ab initio* transcriptome assembly[7,8]. *Ab initio* assembly provides an unbiased modality for gene discovery and has been successful in pinpointing new cancer-associated lncRNAs[9]. Despite such efforts to catalog human lncRNAs, several lines of evidence suggest that the current knowledge of lncRNAs remains inadequate. First, reported discrepancies between independent lncRNA cataloguing efforts suggest that lncRNA annotations are fragmented or incomplete[10]. Second, previous studies largely avoided the annotation of monoexonic transcripts and intragenic lncRNAs owing to the added complexity of transcriptional reconstruction in these regions[11]. Third, the rapid coevolution of high-throughput sequencing technologies and bioinformatics algorithms now enables more accurate transcript reconstruction than was possible with previous efforts[8]. Fourth, high-throughput cataloguing efforts have thus far been confined to select cell lines, individual cancer types or relatively small cohorts[4,9,11]. However, cancers possess highly heterogeneous gene expression patterns, and detecting recurrent expression of subtype-specific lncRNAs will likely require the analysis of much larger tumor cohorts. Here we used a compendium of 7,256 RNA-seq libraries to comprehensively interrogate the human transcriptome, identifying 58,648 lncRNA genes. Moreover, we leveraged our data set to identify a myriad of lncRNAs associated with 27 tissue and cancer types. By uncovering this expansive landscape of tissue- and cancer-associated lncRNAs, we provide the scientific community with a powerful starting point to begin investigating their biological relevance.

[1]Michigan Center for Translational Pathology, University of Michigan, Ann Arbor, Michigan, USA. [2]Department of Computational Medicine and Bioinformatics, Ann Arbor, Michigan, USA. [3]Department of Cellular and Molecular Biology, University of Michigan, Ann Arbor, Michigan, USA. [4]Department of Pathology, University of Michigan, Ann Arbor, Michigan, USA. [5]Howard Hughes Medical Institute, University of Michigan, Ann Arbor, Michigan, USA. [6]Department of Radiation Oncology, University of Michigan, Ann Arbor, Michigan, USA. [7]Section of Thoracic Surgery, Department of Surgery, University of Michigan, Ann Arbor, Michigan, USA. [8]Comprehensive Cancer Center, University of Michigan, Ann Arbor, Michigan, USA. [9]Department of Statistics, Colorado State University, Fort Collins, Colorado, USA. [10]Department of Urology, University of Michigan, Ann Arbor, Michigan, USA. [11]These authors contributed equally to this work. Correspondence should be addressed to A.M.C. (arul@med.umich.edu).

Received 20 June 2014; accepted 18 December 2014; published online 19 January 2015; doi:10.1038/ng.3192

RESULTS

An expanded landscape of human transcription

We attempted to capture the spectrum of human transcriptional diversity by curating 25 independent data sets totaling 7,256 poly(A)$^+$ RNA-seq libraries, including 5,847 from TCGA, 928 from the Michigan Center for Translational Pathology (MCTP), 67 from the Encyclopedia of DNA Elements (ENCODE) and 414 from other public data sets (**Supplementary Fig. 1a** and **Supplementary Tables 1** and **2**). We developed an automated transcriptome assembly pipeline and employed it to process the raw sequencing data sets into *ab initio* transcriptome assemblies (Online Methods, **Supplementary Fig. 1b** and **Supplementary Table 3**). This bioinformatics pipeline used approximately 1,870 core-months (average of 0.26 core-months per library) on high-performance computing environments.

Collectively the RNA-seq data constituted 493 billion fragments; individual libraries averaged 67.9 million total fragments and 55.5 million successful alignments to human chromosomes. On average, 86% of the aligned bases from individual libraries corresponded to annotated RefSeq exons, whereas the remaining 14% fell within introns or intergenic space[12]. We applied coarse quality control measures to account for variations in sequencing throughput, run quality and RNA content by removing 753 libraries with (i) fewer than 20 million total fragments, (ii) fewer than 20 million total aligned reads, (iii) a read length of less than 48 bp or (iv) fewer than 50% of aligned bases corresponding to RefSeq genes (**Supplementary Fig. 1c**,**d**). After coarse filtration, we obtained approximately 391 billion aligned fragments (43.69 Tb of sequence) to use for subsequent analysis. The set of 6,503 libraries passing quality control filters included 6,280 data sets from human tissues and 223 samples from human cell lines. Of the tissue libraries, 5,298 originated from primary tumor specimens, 281 originated from metastases and 701 originated from normal or benign, tumor-adjacent tissues (**Supplementary Fig. 1e**). We subsequently refer to this set of samples as the MiTranscriptome compendium.

To permit sensitive detection of lineage-specific transcription, we partitioned the libraries into 18 cohorts by organ system (**Fig. 1a** and **Supplementary Table 2**) and performed cohort-wise filtering and meta-assembly, before again merging the data (**Fig. 1b**). We developed and employed computational methods to filter out library-specific background noise and predict the most likely isoforms from the assemblies of transcript fragments (transfrags) (**Fig. 1b**). Our filtering approach used transcript abundance and recurrence information to differentiate robust transcription from incompletely processed RNA or contamination from genomic DNA[4] (Online Methods). This stringent approach eliminated the vast majority (>96%) of unannotated transfrags in the compendium (Online Methods and **Supplementary Fig. 2a**–**f**). The remaining transfrags were collapsed into full-length transcript predictions using a greedy dynamic programming algorithm (Online Methods and **Supplementary Fig. 3a**,**b**). For example, in the chromosome 12 locus containing *HOTAIR* and *HOXC11*, the algorithm consolidated 7,471 raw transfrags into 17 transcripts, including ones that accurately matched annotated *HOTAIR* and *HOXC11* isoforms (**Supplementary Fig. 3c**). After merging the meta-assemblies from 18 cohorts for organ systems, we established a consensus set of 384,066 predicted transcripts that we designated as the MiTranscriptome assembly (**Fig. 1b**).

To characterize the MiTranscriptome, we compared it to reference catalogs from RefSeq (December 2013)[12], UCSC (December 2013)[13] and GENCODE (release 19)[10] and to intergenic lncRNA predictions from the previous cataloguing study by Cabili *et al.*[11]. We observed increases in the numbers of exons, splice sites, transcripts and genes

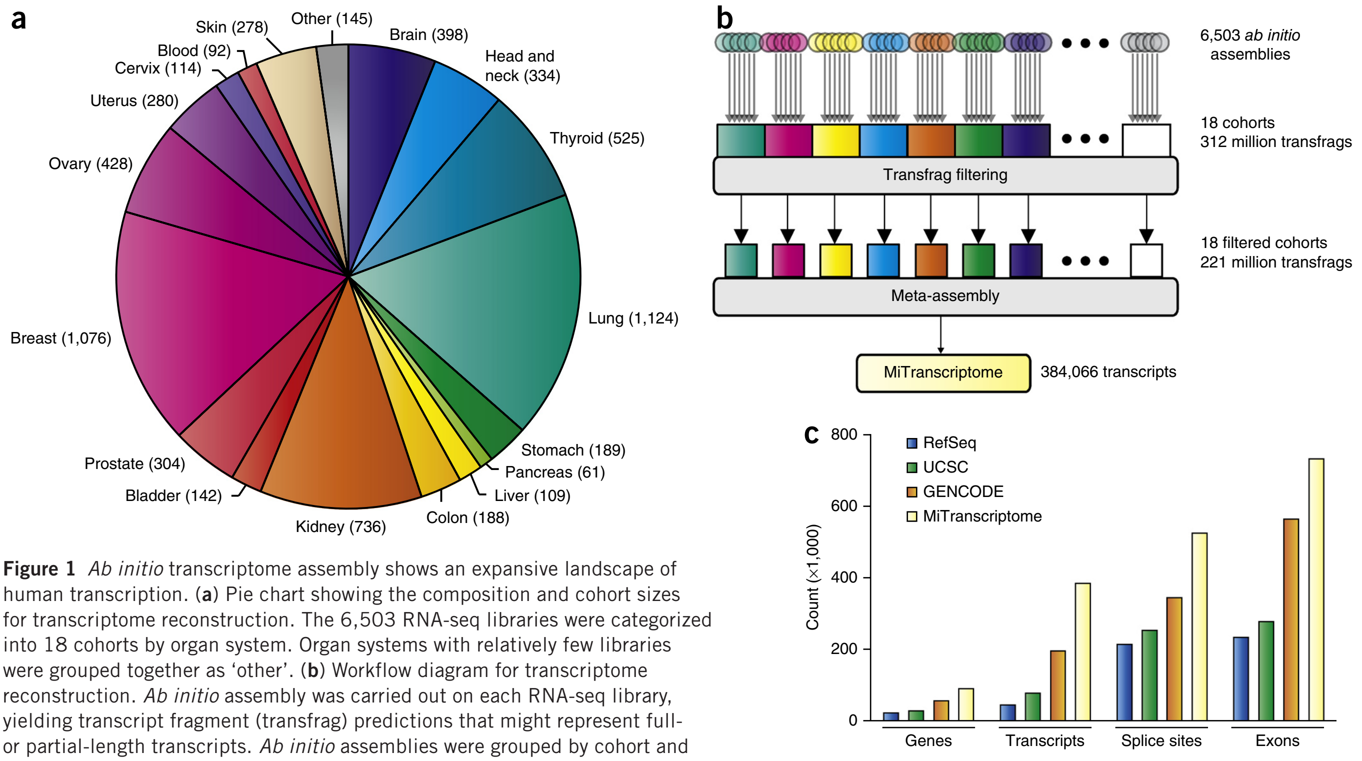


Figure 1 *Ab initio* transcriptome assembly shows an expansive landscape of human transcription. (**a**) Pie chart showing the composition and cohort sizes for transcriptome reconstruction. The 6,503 RNA-seq libraries were categorized into 18 cohorts by organ system. Organ systems with relatively few libraries were grouped together as 'other'. (**b**) Workflow diagram for transcriptome reconstruction. *Ab initio* assembly was carried out on each RNA-seq library, yielding transcript fragment (transfrag) predictions that might represent full- or partial-length transcripts. *Ab initio* assemblies were grouped by cohort and filtered to remove unreliable transfrags. Meta-assembly was performed on the filtered transfrags for each cohort. Finally, the transcripts from the individual cohorts were merged to produce a consensus MiTranscriptome assembly. (**c**) Bar chart comparing the numbers of exons, splice sites, transcripts and genes in the MiTranscriptome assembly with those in the RefSeq (December 2013), UCSC (December 2013) and GENCODE (release 19) catalogs.

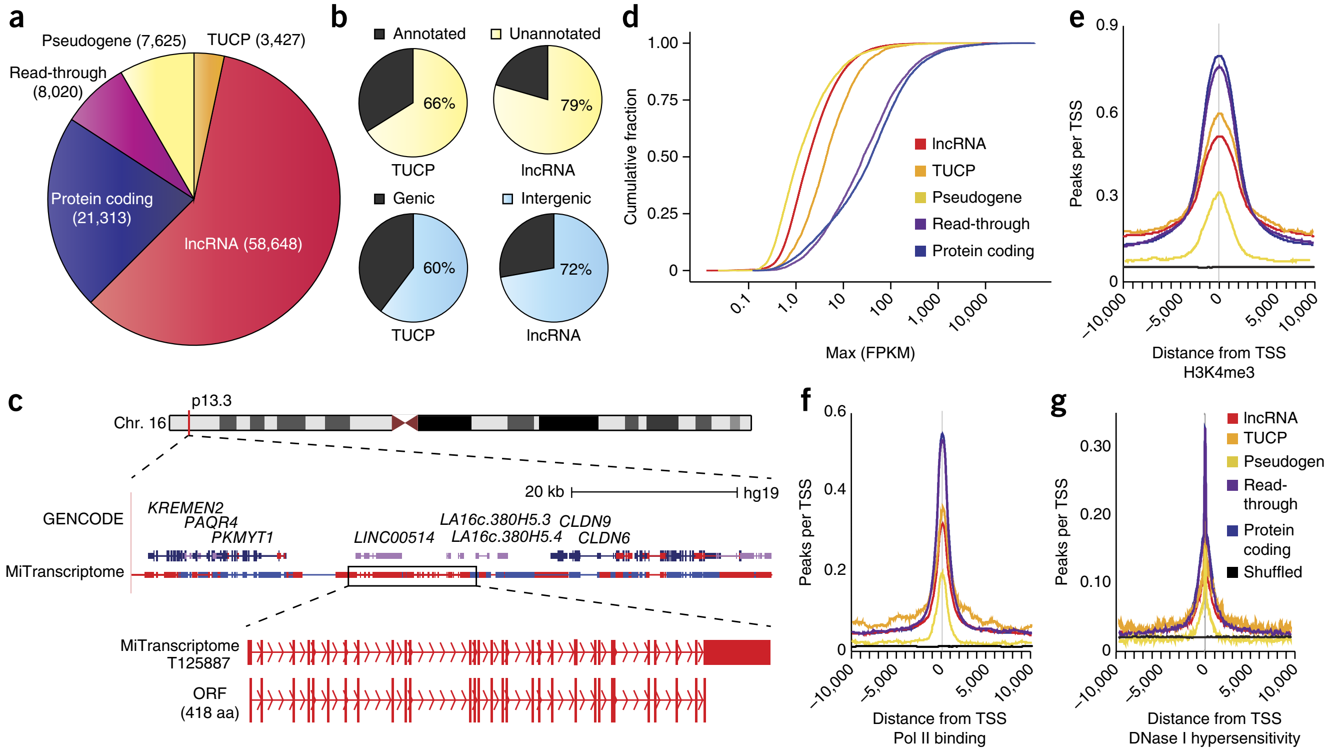


Figure 2 Characterization of the MiTranscriptome assembly. (**a**) Pie chart of the composition and quantities of lncRNA, transcript of unknown coding potential (TUCP), expressed pseudogene, read-through and protein-coding genes in the MiTranscriptome assembly. (**b**) Pie charts of the number of lncRNA and TUCP genes that are unannotated versus annotated relative to reference catalogs (top) and intragenic versus intergenic (bottom). (**c**) Genomic view of the chromosome 16p13.3 locus. Protein-coding genes (*PKMYT1* to *CLDN9*) border an intergenic region containing the GENCODE lncRNA genes *LINC00514* and *LA16c.380H5*. MiTranscriptome transcripts encompassing these genes are shown in a dense view, and an individual isoform containing a 29-exon (418-amino-acid) ORF is highlighted below. This ORF spans multiple GENCODE lncRNAs. (**d**) Empirical cumulative distribution plot comparing the maximum expression (FPKM) of the major isoform of each gene across gene categories. (**e**–**g**) Plots of aggregated ENCODE ChIP-seq data from 13 cell lines at 10-kb intervals surrounding expressed TSSs (FPKM > 0.1) for H3K4me3 (**e**), Pol II binding (**f**) and DNase I hypersensitivity (**g**).

of 29%, 52%, 95% and 57%, respectively, relative to GENCODE, the largest of the reference catalogs (**Fig. 1c** and Online Methods). In terms of well-annotated genes, the assembly demonstrated high sensitivity at the nucleotide and splice-site levels, recovering 94% and 93% of RefSeq nucleotides and splice sites, respectively (**Supplementary Fig. 4a**,**b**). However, detection of precise RefSeq splicing patterns, an ongoing challenge for *in silico* transcriptome reconstruction methods[8], was just 31%. Unannotated transcripts were defined as those lacking strand-specific nucleotide overlap with reference transcripts (RefSeq, UCSC and GENCODE). Although the fraction of transcripts overlapping annotated genes was high in individual cohorts (range of 62–88%, mean of 75%), the fraction of annotated genes within the entire MiTranscriptome was just 46%, alluding to the presence of much unannotated transcription unique to specific lineages (**Supplementary Fig. 4c**).

To assess the robustness of the MiTranscriptome, we stratified transcripts into confidence tiers on the basis of annotation status, the presence of annotated splice junctions, and mono- or multiexonic structure (**Supplementary Table 4**). Using the empirical cumulative distribution function derived from annotated transcript expression levels, we assigned confidence scores to unannotated transcripts (**Supplementary Fig. 5a**). Next, we performed quantitative RT-PCR (qRT-PCR) validations of 100 unannotated transcripts (38 monoexonic and 62 multiexonic) with modest expression (fragments per kilobase of exon per million fragments mapped (FPKM) > 1.0) in at least one of the lung (A549), prostate (LNCaP) and breast (MCF-7) cancer cell lines (Online Methods). To assess false positives arising from background levels of genomic DNA, we also included control reactions without reverse transcriptase. Of the 100 lncRNAs tested, 95 had significantly higher expression in the appropriate cell line than the control (Student's *t* test, $P < 0.05$; **Supplementary Fig. 6**) and showed high correlation between qRT-PCR and RNA-seq expression profiles (**Supplementary Fig. 7a**). In addition, we also performed independent Sanger sequence verification of 18 amplicons that were highly expressed in the 3 cell lines (**Supplementary Fig. 7b**,**c** and **Supplementary Table 5**).

Coding potential assessment of long RNA transcripts

To facilitate further study of the assembly, we classified transcripts into one of five categories: (i) protein coding, (ii) read-through (implying a transcript overlapping multiple separate annotated genes), (iii) pseudogene, (iv) lncRNA and (v) transcript of unknown coding potential (TUCP) (**Supplementary Fig. 8a**). The TUCP classification was originally suggested by Cabili *et al.*[11] and pertains to long RNAs with *in silico* evidence of coding potential. The ability to predict coding potential from sequence features alone has important implications for *ab initio* transcript annotation studies (**Supplementary Note**). Here we predicted TUCPs by incorporating two methods: (i) predictions from the Coding Potential Assessment Tool (CPAT)[14], which analyzes the sequence features of transcript ORFs, and (ii) screening

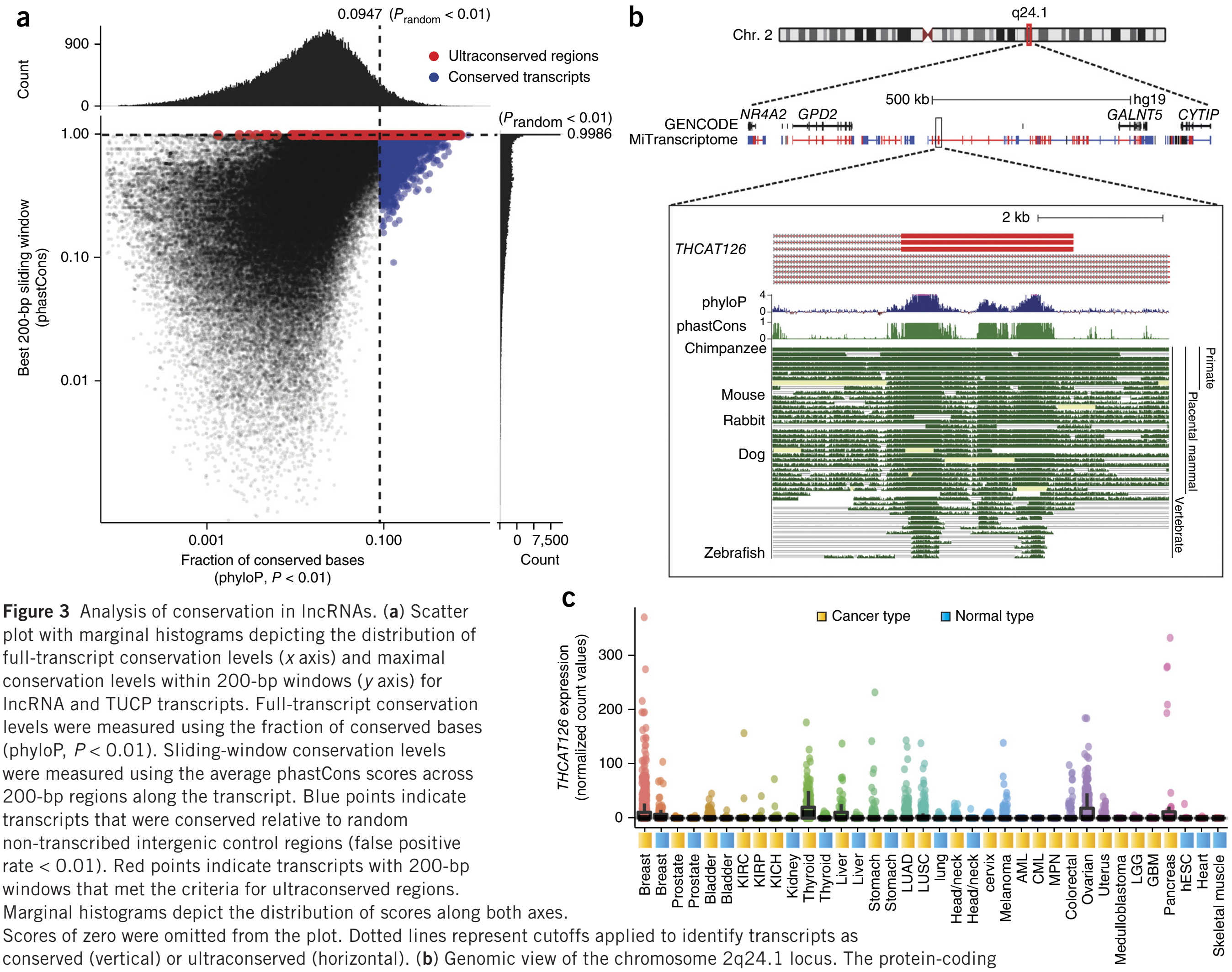


Figure 3 Analysis of conservation in lncRNAs. (**a**) Scatter plot with marginal histograms depicting the distribution of full-transcript conservation levels (*x* axis) and maximal conservation levels within 200-bp windows (*y* axis) for lncRNA and TUCP transcripts. Full-transcript conservation levels were measured using the fraction of conserved bases (phyloP, $P < 0.01$). Sliding-window conservation levels were measured using the average phastCons scores across 200-bp regions along the transcript. Blue points indicate transcripts that were conserved relative to random non-transcribed intergenic control regions (false positive rate < 0.01). Red points indicate transcripts with 200-bp windows that met the criteria for ultraconserved regions. Marginal histograms depict the distribution of scores along both axes. Scores of zero were omitted from the plot. Dotted lines represent cutoffs applied to identify transcripts as conserved (vertical) or ultraconserved (horizontal). (**b**) Genomic view of the chromosome 2q24.1 locus. The protein-coding genes *GALNT5* and *GPD2* flank an intergenic region with no annotated transcripts. MiTranscriptome transcripts are shown in a dense view populating this intergenic space. Blue and red color represent positive- and negative-strand transcripts, respectively (this color scheme applies to all subsequent genomic views). The most magnified view at the bottom depicts a highly conserved exon from the lncRNA *THCAT126*. The Multiz alignment of 46 vertebrate species is depicted as well as the per-base phyloP and phastCons conservation scores. (**c**) Expression data for *THCAT126* across all MiTranscriptome cancer and normal tissue type cohorts.

for the presence of a known Pfam domain[15] within a transcript ORF (**Supplementary Fig. 8b–h** and **Supplementary Note**).

Remarkably, over 60% of MiTranscriptome genes were classified as either lncRNAs or TUCPs (59% lncRNAs and 3.5% TUCPs; **Fig. 2a**). The majority of lncRNAs and TUCPs were unannotated relative to RefSeq, UCSC and GENCODE genes (79% and 66%, respectively) and were located within intergenic regions (72% and 60%, respectively) (**Fig. 2b**). Interestingly, 5,248 transcripts overlapping annotated lncRNAs were flagged as TUCPs, suggesting that previous annotation attempts identified ostensibly noncoding fragments of transcripts possessing robust ORFs. For example, in a chromosome 16 intergenic locus, we detected transcripts harboring an ORF, predicted to encode a 418-amino-acid product, containing 29 exons that overlapped 3 independent genes annotated by GENCODE as lncRNAs (*LINC00514*, *LA16c-380H5.3* and *LA16c-380H5.4*), suggesting that some annotated lncRNAs might in fact be inaccurate partial representations of a larger protein-coding gene (**Fig. 2c**). To further investigate coding potential, we searched a large human proteomics data set derived from benign tissue samples[16] for peptides uniquely mapping to TUCP ORFs and noted 268 such genes (**Supplementary Table 6**). Given these intriguing results, we anticipate that future integration of proteomics data from tumor tissues will strengthen our TUCP predictions.

Characterization of long RNAs

lncRNA and TUCP genes tended to have fewer exons than read-through or protein-coding genes, but we nevertheless observed appreciable alternative splicing for all classes of transcripts[11,17] (**Supplementary Fig. 5b**). Furthermore, we observed that lncRNAs and TUCPs were expressed at lower levels than read-through or protein-coding transcripts, which is consistent with previous studies[9,11,17,18] (**Fig. 2d**). To further corroborate active transcription of the lncRNAs and TUCPs, we intersected intervals surrounding the transcription start sites (TSSs) with ENCODE chromatin immunoprecipitation and sequencing (ChIP-seq) data for histone 3 lysine 4 trimethylation (H3K4me3), RNA polymerase II (Pol II) binding sites and DNase I hypersensitivity data from 13 cell lines[19,20] (Online Methods). Maximal enrichment of these marks at the TSSs of these genes but not at randomly shuffled control regions suggests that the assembled lncRNA and TUCP transcripts possess actively regulated promoters (**Fig. 2e**–**g**).

lncRNAs harboring conserved elements

The evolutionary conservation of lncRNAs has been a topic of ongoing conversation, with several reports suggesting that lncRNAs are modestly conserved[11,17,18,21]. In agreement with previous reports, we observed increases in both transcript and promoter conservation levels for lncRNAs and TUCPs relative to random control regions (Online Methods and **Supplementary Fig. 5c**–**f**). Shifts in the cumulative distributions of lncRNA and TUCP transcripts were greater for annotated transcripts than for unannotated transcripts. This difference might reflect discovery bias toward highly conserved genes detectable across multiple model systems. Moreover, the subtle increases in conservation we observe for lncRNAs suggest, at least in humans, that lncRNA conservation might be an exceptional phenomenon rather than a general one. Therefore, we specifically delineated 3,309 lncRNAs (5.6% of all lncRNAs) harboring markedly higher base-wise conservation than random intergenic regions to enable the focused study of these transcripts (**Fig. 3a**, Online Methods and **Supplementary Fig. 5e**). In addition, an intriguing aspect of the noncoding genome includes ultraconserved elements (UCEs), which are stretches of DNA >200 nt in length with nearly perfect sequence identity across multiple organisms[22,23]. We delineated 597 intergenic lncRNAs (1.2% of all intergenic lncRNAs) harboring UCEs and designated these as highly conserved long intergenic noncoding RNAs (HICLINCs) (Online Methods and **Supplementary Fig. 5h**). For example, *THCAT126*, a previously unannotated intergenic lncRNA on chromosome 2q24, contains elements in its final exons that are conserved in nearly all vertebrates including zebrafish (**Fig. 3b**). Moreover, *THCAT126* is expressed widely across many tissue types, including thyroid cancer (**Fig. 3c**). Highly conserved lncRNAs such as *THCAT126* (and other cancer-associated HICLINCs described below) provide an exciting avenue for *in vivo* study of the role of lncRNAs in development and carcinogenesis.

lncRNAs overlapping disease-associated SNPs

To investigate the relationship of the MiTranscriptome assembly with disease-associated regions of the genome, we assessed the overlap of transcripts in the assembly with 11,194 unique disease-associated SNPs from a catalog of genome-wide association studies (GWAS)[24]. MiTranscriptome exons and transcripts overlapped 2,586 and 9,770 GWAS SNPs in comparison to just 1,096 and 7,050 SNPs overlapped by reference transcripts, respectively (**Supplementary Fig. 9a**,**b**). Altogether, transcripts in the assembly overlapped 2,881 formerly intergenic SNPs located within gene deserts and only lacked 161 GWAS SNPs overlapping annotated genes. We tested for the possibility that the increased overlap with GWAS SNPs occurred at a rate above chance and observed that both MiTranscriptome transcripts and exons were significantly enriched for GWAS SNPs relative to random SNPs chosen from the same chip platform (paired *t* test, $P = 5.25 \times 10^{-135}$ and 1.15×10^{-199}, respectively; Online Methods, **Supplementary Fig. 9c** and **Supplementary Note**). Moreover, unannotated intergenic lncRNAs and TUCPs were also significantly enriched for disease-associated regions, with exons more highly enriched than full-length transcripts (paired *t* test, $P = 9.90 \times 10^{-78}$ and 5.50×10^{-50} for whole transcripts and exons, respectively; **Supplementary Fig. 9d**). These data argue that a rigorous reevaluation of the regulation of allele-specific gene expression in regions proximal to GWAS SNPs might yield informative biological associations with the new lncRNAs identified in this study.

Differential expression analysis

Our large-scale transcriptome reconstruction process unveiled tremendous transcriptional complexity highlighted by the presence of thousands of uncharacterized lncRNAs and TUCPs. To prioritize disease-associated and lineage-specific transcription, we developed a non-parametric method for the testing of differential expression called Sample Set Enrichment Analysis (SSEA) (Online Methods and **Supplementary Note**). SSEA adapts the weighted Kolmorgorov-Smirnoff–like tests used by Gene Set Enrichment Analysis (GSEA)[25] to discover transcript expression changes between two groups of samples. The non-parametric nature of this method permits sensitive detection of differential expression within heterogeneous sample populations (for example, tumor subtypes). We performed 50 analyses of differential expression including various cancer or normal lineage types (one cancer or lineage type versus all other MiTranscriptome samples) and cancer versus normal comparisons within a single tissue type (**Fig. 4a** and Online Methods). Collectively, SSEA detected over 2 million significant associations (false discovery rate (FDR) $< 1 \times 10^{-3}$ for cancer versus normal analyses and FDR $< 1 \times 10^{-7}$ for lineage analyses) involving 267,726 MiTranscriptome transcripts (Online Methods and **Supplementary Table 7**). To validate the enrichment testing approach, we assessed its ability to rediscover known biomarkers upregulated and downregulated in prostate and breast cancers. We assessed the concordance between the top 1% of positively and negatively enriched genes from each cancer type with cancer gene signatures obtained from the Oncomine database of microarray studies[26–32] (Online Methods and **Supplementary Table 8**). A heat map of the odds ratios of the gene signature associations showed striking agreement between SSEA and the other studies, with SSEA often demonstrating equal or better concordance to each microarray study than comparison between microarray studies (**Fig. 4b** and **Supplementary Table 9**). Thus, testing for isoform-level differential expression from the MiTranscriptome *ab initio* assembly of RNA-seq data recapitulated the results from cancer microarray gene expression studies, supporting the SSEA method as a viable tool for the detection of differential expression.

To further credential the enrichment testing approach, we assessed its ability to detect positive control lncRNAs and protein-coding genes in breast and prostate cancers. For example, SSEA correctly identified the oncogenic lncRNA *HOTAIR*, *ESR1* (encoding estrogen receptor 1) and *GATA3* (encoding GATA-binding protein 3) as highly positively enriched in breast cancers and accurately nominated the tumor-suppressor lncRNA *MEG3* and the metastasis-suppressor *LIFR*[33] as highly negatively enriched[30,31,34,35] (**Fig. 4c**–**e**). Similarly, in prostate cancers, SSEA detected differential expression of lncRNAs and protein-coding genes consistent with the literature (**Fig. 4f**). Notably, the known prostate cancer lncRNAs *PCA3* (prostate cancer antigen-3) and *SChLAP1* were strikingly enriched in a cancer-specific and prostate-specific manner relative to all other sample set analyses (**Fig. 4g**,**h**)[28,36]. Overall, the ability of the enrichment testing approach to rediscover known cancer-associated genes in an unbiased fashion indicates its usefulness for the analysis of cancer association and lineage specificity within the panorama of uncharacterized transcription unveiled by MiTranscriptome.

Characterization of differentially expressed lncRNAs

To extend our study beyond known cancer-associated genes, we mined the enrichment test results for lineage-specific and cancer-specific transcripts in an unbiased manner. Lineage specificity was assayed using sample sets for each cancer or tissue type in comparison to all other samples in the MiTranscriptome compendium (**Fig. 4a**, "Cancer types" and "Normal types"), and SSEA results were used

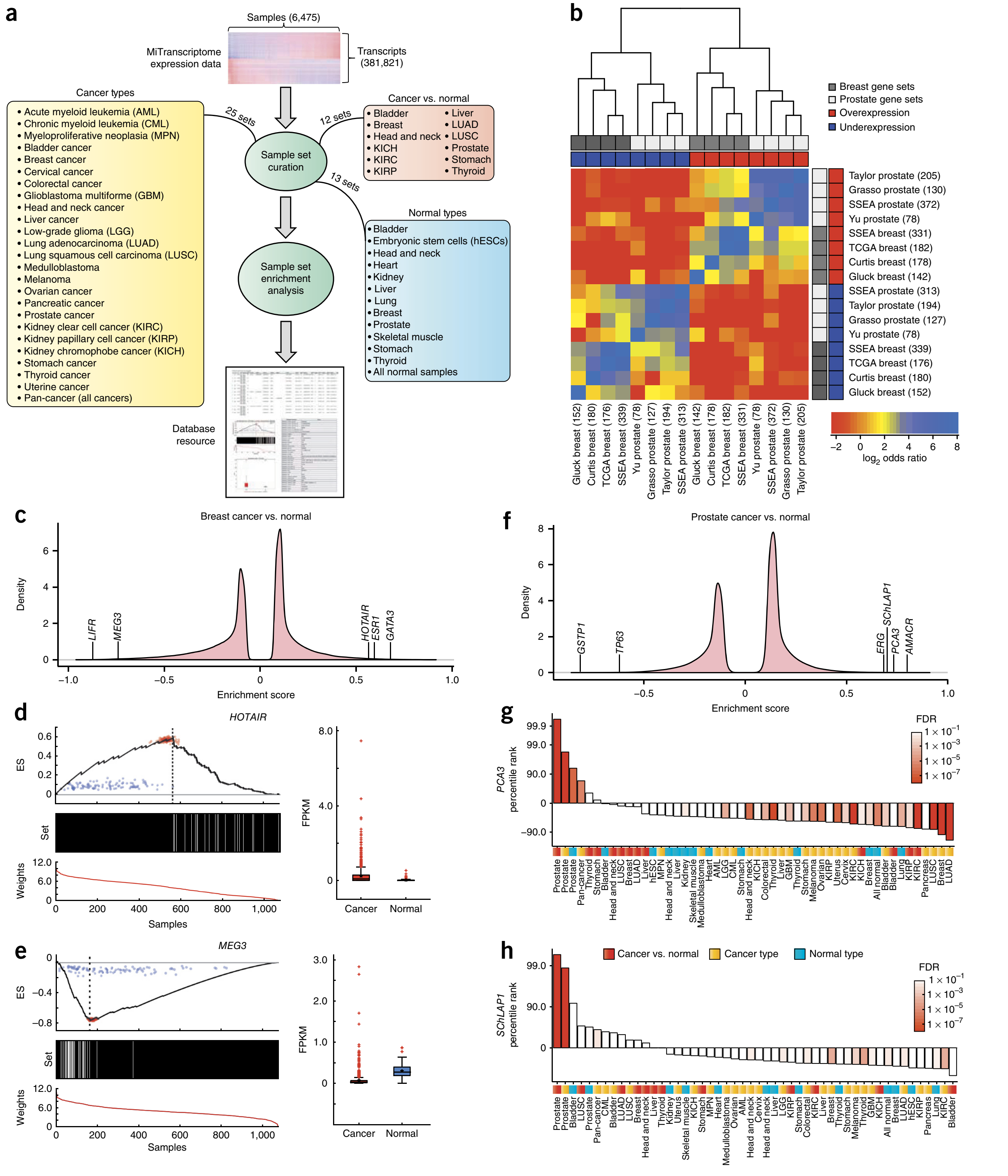


Figure 4 Methodology for discovering cancer-associated lncRNAs. (**a**) Samples were grouped into 50 different sample sets in 3 categories: (i) cancer types, (ii) normal types and (iii) cancer versus normal. Enrichment testing was performed using SSEA, and significant transcripts were imported into an online resource. (**b**) Heat map showing the concordance of the SSEA algorithm with the prostate and breast cancer gene signatures obtained from the Oncomine database. The top 1% of overexpressed and underexpressed genes from each analysis were compared using Fisher's exact tests. (**c**) Enrichment score density plots for breast cancers versus normal samples. (**d**,**e**) Enrichment and expression plots for the lncRNAs *HOTAIR* (**d**) and *MEG3* (**e**). Subplots include the running enrichment score (ES) across all samples (dotted lines, maximum and minimum enrichment scores; red points, Poisson resamplings of fragment counts; blue points, random permutations of the sample labels) (top); identity as a cancer (black bars) or normal (white bars) sample (middle); and rank-ordered normalized expression values (bottom). Adjacent box plots (interquartile range and median shown by box and whiskers) depict transcript expression (FPKM) in cancer and normal samples (967 and 109 patients in the breast cancer and normal groups, respectively). (**f**) Enrichment score density plots for prostate cancer versus normal samples. (**g**,**h**) Bar plots of the percentile ranks for the prostate cancer–specific lncRNAs *PCA3* (**g**) and *SChLAP1* (**h**) across the cancer versus normal (red), cancer type (gold) and normal type (blue) sample sets. Bar colors depict statistical significance (FDR).

to determine the degree of enrichment for each transcript in the various cancer and tissue types. Unsupervised clustering of transcript percentile ranks for the top 1% of transcripts in each lineage demonstrated distinct lineage-specific signatures while also suggesting relationships among lineages and between cancer and normal sets from the same lineage (Online Methods and **Supplementary Fig. 10a**). Examples of closely related lineage clusters include blood cancers (acute myeloid leukemia (AML), chronic

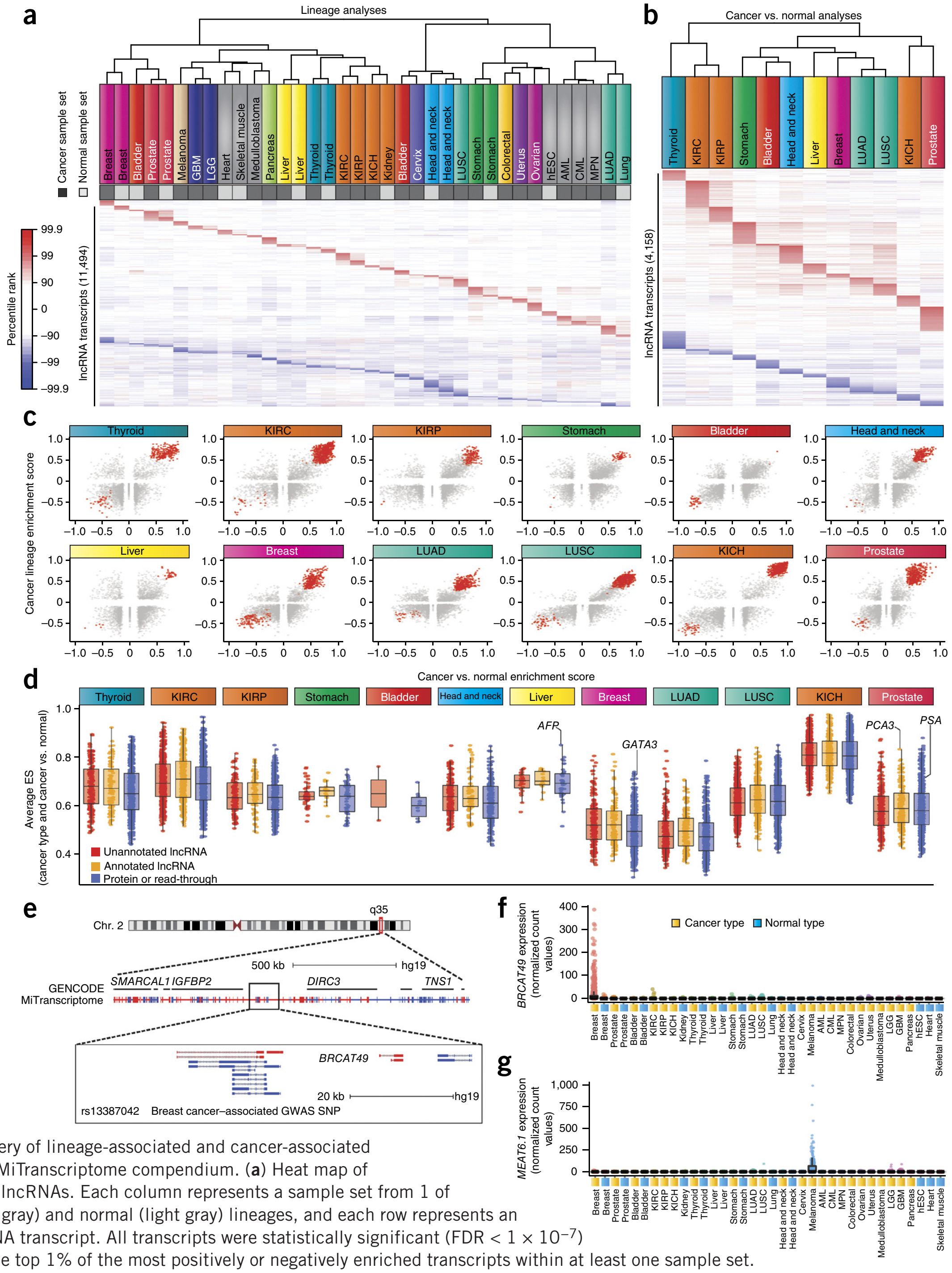


Figure 5 Discovery of lineage-associated and cancer-associated lncRNAs in the MiTranscriptome compendium. (**a**) Heat map of lineage-specific lncRNAs. Each column represents a sample set from 1 of 25 cancer (dark gray) and normal (light gray) lineages, and each row represents an individual lncRNA transcript. All transcripts were statistically significant (FDR $< 1 \times 10^{-7}$) and ranked in the top 1% of the most positively or negatively enriched transcripts within at least one sample set. The heat map color spectrum corresponds to percentile ranks, with underexpressed transcripts (blue) and overexpressed transcripts (red). (**b**) Heat map of cancer-specific lncRNAs nominated by SSEA cancer versus normal analysis of 12 cancer types (columns). All transcripts were statistically significant (FDR $< 1 \times 10^{-3}$) and ranked in the top 1% of the most positively or negatively enriched transcripts within at least one sample set. (**c**) Scatter plots showing enrichment score for cancer versus normal (*x* axis) and cancer lineage (*y* axis) for all lineage-specific and cancer-associated lncRNA transcripts across 12 cancer types. Red points indicate transcripts meeting the percentile cutoffs for cancer and lineage association. (**d**) Box plot comparing the performance of cancer- and lineage-associated lncRNAs across 12 cancer types. The average of the lineage and cancer versus normal enrichment scores is plotted on the *y* axis. (**e**) Genomic view of the chromosome 2q35 locus. The most magnified view at the bottom depicts *BRCAT49*, a breast lineage– and breast cancer–specific lncRNA. The breast cancer–associated GWAS SNP rs13387042 is depicted in green. (**f**) Expression data for *BRCAT49* across all MiTranscriptome cancer and normal tissue type cohorts. (**g**) Expression data for *MEAT6* across all MiTranscriptome cancer and normal tissue type cohorts.

Table 1 Summary of lineage- and/or cancer-specific lncRNAs nominated in this study

Tissue or cancer type (naming convention)	Total number of associated noncoding transcripts	Number of cancer- and tissue-specific transcripts	Number of conserved transcripts	Number of transcripts containing UCEs	Number of transcripts classified as TUCPs
Acute myelogenous leukemia–associated transcripts (AMATs)	373	NA	29	13	26
Bladder cancer–associated transcripts (BLCATs)	61	0	9	2	5
Breast cancer–associated transcripts (BRCATs)	1,115	134	82	27	76
Cervical cancer–associated transcripts (CVATs)	162	NA	12	2	13
Chronic myelogenous leukemia–associated transcripts (CMATs)	157	NA	16	3	11
Colorectal cancer–associated transcripts (CRATs)	163	NA	29	4	17
Glioblastoma multiforme–associated transcripts (GBATs)	161	NA	11	2	22
Head and neck cancer–associated transcripts (HNCATs)	766	5	45	15	68
Heart tissue–associated transcripts (HRATs)	170	NA	16	1	12
Human embryonic stem cell–associated transcripts (ESATs)	205	NA	10	0	20
Chromophobe renal cell carcinoma–associated transcripts (KCHCATs)	1,050	52	64	20	92
Renal clear cell carcinoma–associated transcripts (KCCATs)	1,429	215	84	26	123
Renal papillary cell carcinoma–associated transcripts (KPCATs)	474	0	41	8	38
Low-grade glioma–associated transcripts (LGATs)	265	NA	31	10	23
Liver cancer–associated transcripts (LVCATs)	250	0	18	1	20
Lung adenocarcinoma–associated transcripts (LACATs)	953	19	64	19	61
Lung squamous cell carcinoma–associated transcripts (LSCATs)	1,014	10	70	23	58
Medulloblastoma–associated transcripts (MBATs)	312	NA	26	3	33
Melanoma-associated transcripts (MEATs)	339	NA	24	2	34
Myeloproliferative neoplasia–associated transcripts (MPATs)	101	NA	12	1	8
Ovarian cancer–associated transcripts (OVATs)	163	NA	37	12	30
Pancreatic cancer–associated transcripts (PNATs)	247	NA	27	4	22
Prostate cancer–associated transcripts (PRCATs)	727	38	49	14	62
Skeletal muscle tissue–associated transcripts (SMATs)	123	NA	5	1	11
Stomach cancer–associated transcripts (STCATs)	95	0	10	1	10
Thyroid cancer–associated transcripts (THCATs)	1,289	80	73	21	111
Uterine endometrial carcinoma–associated transcripts (UTATs)	183	NA	31	1	16

For each of the 27 tissue types (rows), the table lists the numbers of lncRNA genes associated with the tissue and/or cancer type, enriched for conserved nucleotides, containing UCEs and classified as TUCPs. Cancer- and tissue-specific lncRNAs are only delineated for tissue types for which there was a sufficient number of matched normal samples to perform a cancer versus normal analysis (NA reported otherwise).

myeloid leukemia (CML) and myeloproliferative neoplasia (MPN)), brain cancers (low-grade glioma (LGG) and glioblastoma multiforme (GBM)) and muscle tissue (cardiac and skeletal). Additionally, a cluster comprising cervical cancer, head and neck cancer and normal lineages, lung squamous cell cancer and bladder cancer emerged, suggesting that primarily squamous (and transitional) cell carcinomas from distant primary sites share important gene expression relationships. Intriguingly, unsupervised clustering of only the lncRNAs in the top 1% of the SSEA analysis for lineage association recapitulated all of these relationships, indicating the capacity for lncRNAs to independently identify cancer and normal lineages (**Fig. 5a**).

Next, we investigated the dimension of cancer-specific transcriptional dynamics in 12 tissues with ample numbers of both cancer and normal samples (**Fig. 4a**, "Cancer versus normal"). Similar to above, unsupervised clustering of the top 1% of cancer-associated lncRNAs demonstrated highly specific signatures for each cancer type, with the exception of lung and kidney cancers (**Fig. 5b** and **Supplementary Fig. 10b**). Lung squamous cell carcinomas (LUSC) and adenocarcinomas (LUAD) clustered together and shared numerous transcripts with cancer association. Similarly, renal clear cell (KIRC) and papillary cell (KIRP) carcinomas exhibited highly overlapping signatures, whereas renal chromophobe carcinomas (KICH) remained distinct from KIRC and KIRP.

Finally, we intersected the results from the lineage and cancer analyses. With extensive further evaluation, such transcripts might have translational potential for use in non-invasive clinical tests, particularly for cancers that lack reliable biomarkers. Notable examples included the prostate-specific lncRNAs *PCA3* and *SChLAP1* presented earlier (**Fig. 4g**,**h**). A myriad of lncRNAs were detected as being lineage and cancer associated (in the top 5% of both analyses) for each of the cancer types analyzed (**Fig. 5c** and **Supplementary Fig. 11a**). A direct comparison of lncRNAs and protein-coding transcripts showed that both annotated and unannotated lncRNAs have the potential to perform at a comparable level to protein-coding genes, supporting a role for lncRNAs in augmenting tissue- and cancer-specific gene signatures (**Fig. 5d** and **Supplementary Fig. 11b**,**c**).

We applied stringent statistical cutoffs to nominate 7,942 lncRNA or TUCP genes (11,478 transcripts) as cancer associated, lineage associated or both (Online Methods and **Supplementary Table 10**). Transcripts meeting the stringent cutoffs in the cancer versus normal analyses were designated as having cancer association. Those transcripts meeting stringent cutoffs for lineage specificity in noncancerous tissue (for example, heart, skeletal muscle or embryonic stem cells) and in cancers lacking RNA-seq data for benign tissue were designated as lineage associated. Moreover, transcripts meeting the cutoffs for both the cancer versus normal and lineage specificity analyses were designated as having cancer and lineage association (**Table 1**). Transcripts with significant association in just one tissue type were given names according to that tissue type (**Table 1**), and transcripts with associations in multiple tissues were named cancer-associated transcripts (CATs). An additional 545 lncRNA genes (1,634 transcripts) that possessed UCEs but did not meet the stringent lineage and cancer association criteria were designated as HICLINCs. Of these 8,487 lncRNAs, 7,804 did not possess an official gene symbol according to the Human Genome Organization (HUGO) Gene Nomenclature Committee[33] and were thus named according to the convention described in **Table 1**.

To infer putative roles for cancer- or lineage-associated lncRNAs in oncogenesis, we curated 2,078 MSigDB gene sets into categories corresponding to biological function (angiogenesis and hypoxia, metastasis, proliferation and cell cycle, cell adhesion, and DNA damage and repair) or signatures from gene expression profiling studies (**Supplementary Table 11**)[25]. We constructed an expression correlation matrix between lncRNAs and protein-coding genes and employed a 'guilt-by-association' analysis whereby the correlation data were processed by GSEA to generate a matrix of the association of each lncRNA with each gene set, capturing over 14,000 transcripts with significant associations (family-wise error rate (FWER) < 0.001; Online Methods and **Supplementary Tables 12** and **13**)[37].

To allow the scientific community to explore our discoveries, we developed an online portal featuring detailed characteristics of the nominated transcripts (see URLs) and present several examples of intriguing lncRNAs here. First, *BRCAT49* (breast cancer–associated transcript-49) is a breast cancer– and lineage-associated lncRNA gene (**Fig. 5d**) located ~45 kb downstream of the intergenic breast cancer–associated SNP rs13387042 that has been implicated by multiple GWAS (**Fig. 5e**,**f**)[38–42]. *BRCAT49* provides a possible target for explaining the breast cancer association of this genomic region and would be a candidate for intergenic expression quantitative trait locus (eQTL) analysis. We also performed further interrogation of the relationship with GWAS SNPs, and all transcripts within 50 kb of a GWAS SNP implicated in a disease locus for which the lncRNA was identified as having a significant association are reported in **Supplementary Table 14**. Second, the lncRNA gene we termed *MEAT6* (melanoma-associated transcript-6) was found to be in the 99.8th percentile in the melanoma lineage SSEA analysis (**Fig. 5a**). Genomic investigation delineated *MEAT6* as a partially annotated transcriptional variant of the lncRNA *AK090788* on chromosome 6q26 (**Supplementary Fig. 12a**). However, *MEAT6* uses an alternative start site and upstream exons absent from reference catalogs. Expression of *MEAT6* isoforms using the novel start site was highly specific to the melanoma samples in the MiTranscriptome cohort (**Fig. 5g**); in contrast, isoforms lacking the *MEAT6* start site had a dramatically different pan-cancer expression profile with almost no expression in melanoma (**Supplementary Fig. 12b**). Additional examples of expression profiles for cancer- or lineage-specific lncRNAs in other tissue types are displayed in **Supplementary Figure 12c**,**d**. The examples shown here are indeed representative, and we anticipate that an abundance of uncharacterized transcription with biological and translational potential can be leveraged using our discoveries here and our online resource (see URLs; **Supplementary Tables 10** and **11**).

DISCUSSION

Here we discovered and characterized an expanded landscape of transcription via unbiased transcriptome reconstruction from thousands of tumors, normal tissues and cell lines. Our work uses several orders of magnitude more RNA-seq data (~100-fold) than previous RNA-seq lncRNA discovery efforts and vastly increases the universe of known transcripts in both normal tissues and cancer. The unprecedented breadth (6,503 samples) and depth (>43 Tb of sequence) of our compendium enabled the sensitive detection of robust transcription and the specific filtering out of background noise. The lncRNAs in our assembly (58,648 genes, often with multiple isoforms) far outnumber entries in current lncRNA databases (<16,000 genes), implying that reference transcript annotations might be fragmented or otherwise incomplete[11,17,43–46]. Moreover, our assembly indicates that the genomic diversity of lncRNAs eclipses that of coding transcripts (with nearly 60,000 lncRNA genes versus approximately 30,000 protein-coding genes), a disparity that may grow as additional diseases and cell types are sequenced and more lncRNAs are discovered.

Multiple lines of *in silico* evidence support the biological and functional relevance of MiTranscriptome transcripts, including robust expression, protein-coding potential (for TUCPs), high conservation, active regulation at promoters, proximity to disease-associated genomic polymorphisms, correlation with protein-coding gene signatures, lineage specificity and cancer specificity. Moreover, many lncRNAs independently identified by this study have previously been validated and mechanistically linked to carcinogenesis (**Supplementary Table 15**)[35,36,47–49]. Regardless of their functional contributions, uncharacterized MiTranscriptome transcripts could serve as future cancer biomarkers.

Although the central dogma remains a core tenet of cellular and molecular biology, the appreciation of lncRNAs as functional genomic elements that defy the central dogma may be essential for fully understanding biology and disease. Taken together, our results indicate that the vastness and complexity of lncRNA transcription has been grossly underappreciated and that a myriad of lncRNAs are associated with carcinogenesis. We anticipate that the MiTranscriptome assembly and lncRNAs identified by this study, as well as the computational tools developed herein, will provide a foundation for lncRNA genomics, biomarker development and the delineation of cancer disease mechanisms.

URLs. MiTranscriptome Online Portal, http://mitranscriptome.org/.

METHODS

Methods and any associated references are available in the online version of the paper.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS

We thank B. Palen and J. Hallum for technical assistance with the high-performance computing cluster, S. Roychowdhury for reviewing the manuscript, the University of Michigan DNA Sequencing Core for Sanger sequencing and K. Giles for critically reading the manuscript and for the submission of documents. This work was supported in part by US National Institutes of Health Prostate Specialized Program of Research Excellence grant P50 CA69568, Early Detection Research Network grant UO1 CA111275, US National Institutes of Health grants R01 CA132874 and RO1 CA154365 (D.G.B. and A.M.C.), and US Department of Defense grant PC100171 (A.M.C.). A.M.C. is supported by the Prostate Cancer Foundation and the Howard Hughes Medical Institute. A.M.C. is an American Cancer Society Research Professor and a Taubman Scholar of the University of Michigan. R.M. was supported by a Prostate Cancer Foundation Young Investigator Award and by US Department of Defense Post-Doctoral Fellowship W81XWH-13-1-0284. Y.S.N. is supported by a University of Michigan Cellular and Molecular Biology National Research Service Award Institutional Predoctoral Training Grant.

AUTHOR CONTRIBUTIONS

M.K.I., Y.S.N. and A.M.C. conceived the study and analyses. M.K.I. processed RNA-seq data and performed *ab initio* assembly. M.K.I. and Y.S.N. performed data processing and data analysis with assistance from T.R.B., R.M., A.S., Y.H., J.R.E., S.Z., J.R.P. and F.Y.F. R.M., U.S., A.S. and Y.H. performed quantitative PCR validations. M.K.I. and Y.S.N. developed SSEA with the help of H.K.I. D.G.B. contributed primary samples. D.R.R., Y.-M.W. and S.M.D. generated RNA-seq libraries, and X.C. performed the sequencing. M.K.I., Y.S.N. and A.S. developed the web resource. T.R.B. provided systems administration, data storage,

high-performance computing and networking support. A.P. performed the proteomics analysis. M.K.I., Y.S.N. and A.M.C. wrote the manuscript. All authors discussed results and commented on the manuscript.

COMPETING FINANCIAL INTERESTS

The authors declare competing financial interests: details are available in the online version of the paper.

Reprints and permissions information is available online at http://www.nature.com/reprints/index.html.

1. Ferlay, J. *et al.* Cancer incidence and mortality worldwide: sources, methods and major patterns in GLOBOCAN 2012. *Int. J. Cancer* **136**, E359–E386 (2015).
2. Kandoth, C. *et al.* Mutational landscape and significance across 12 major cancer types. *Nature* **502**, 333–339 (2013).
3. Ciriello, G. *et al.* Emerging landscape of oncogenic signatures across human cancers. *Nat. Genet.* **45**, 1127–1133 (2013).
4. Djebali, S. *et al.* Landscape of transcription in human cells. *Nature* **489**, 101–108 (2012).
5. Ulitsky, I. & Bartel, D.P. lincRNAs: genomics, evolution, and mechanisms. *Cell* **154**, 26–46 (2013).
6. Prensner, J.R. & Chinnaiyan, A.M. The emergence of lncRNAs in cancer biology. *Cancer Discov.* **1**, 391–407 (2011).
7. Trapnell, C. *et al.* Differential analysis of gene regulation at transcript resolution with RNA-seq. *Nat. Biotechnol.* **31**, 46–53 (2013).
8. Steijger, T. *et al.* Assessment of transcript reconstruction methods for RNA-seq. *Nat. Methods* **10**, 1177–1184 (2013).
9. Prensner, J.R. *et al.* Transcriptome sequencing across a prostate cancer cohort identifies *PCAT-1*, an unannotated lincRNA implicated in disease progression. *Nat. Biotechnol.* **29**, 742–749 (2011).
10. Harrow, J. *et al.* GENCODE: the reference human genome annotation for The ENCODE Project. *Genome Res.* **22**, 1760–1774 (2012).
11. Cabili, M.N. *et al.* Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. *Genes Dev.* **25**, 1915–1927 (2011).
12. Pruitt, K.D. *et al.* RefSeq: an update on mammalian reference sequences. *Nucleic Acids Res.* **42**, D756–D763 (2014).
13. Karolchik, D. *et al.* The UCSC Genome Browser database: 2014 update. *Nucleic Acids Res.* **42**, D764–D770 (2014).
14. Wang, L. *et al.* CPAT: Coding-Potential Assessment Tool using an alignment-free logistic regression model. *Nucleic Acids Res.* **41**, e74 (2013).
15. Finn, R.D. *et al.* Pfam: the protein families database. *Nucleic Acids Res.* **42**, D222–D230 (2014).
16. Kim, M.S. *et al.* A draft map of the human proteome. *Nature* **509**, 575–581 (2014).
17. Derrien, T. *et al.* The GENCODE v7 catalog of human long noncoding RNAs: analysis of their gene structure, evolution, and expression. *Genome Res.* **22**, 1775–1789 (2012).
18. Guttman, M. *et al.* *Ab initio* reconstruction of cell type–specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincRNAs. *Nat. Biotechnol.* **28**, 503–510 (2010).
19. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012).
20. Rosenbloom, K.R. *et al.* ENCODE data in the UCSC genome browser: year 5 update. *Nucleic Acids Res.* **41**, D56–D63 (2013).
21. Necsulea, A. *et al.* The evolution of lncRNA repertoires and expression patterns in tetrapods. *Nature* **505**, 635–640 (2014).
22. Dimitrieva, S. & Bucher, P. UCNEbase—a database of ultraconserved non-coding elements and genomic regulatory blocks. *Nucleic Acids Res.* **41**, D101–D109 (2013).
23. Bejerano, G. *et al.* Ultraconserved elements in the human genome. *Science* **304**, 1321–1325 (2004).
24. Welter, D. *et al.* The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic Acids Res.* **42**, D1001–D1006 (2014).
25. Subramanian, A. *et al.* Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. USA* **102**, 15545–15550 (2005).
26. Grasso, C.S. *et al.* The mutational landscape of lethal castration-resistant prostate cancer. *Nature* **487**, 239–243 (2012).
27. Yu, Y.P. *et al.* Gene expression alterations in prostate cancer predicting tumor aggression and preceding development of malignancy. *J. Clin. Oncol.* **22**, 2790–2799 (2004).
28. Taylor, B.S. *et al.* Integrative genomic profiling of human prostate cancer. *Cancer Cell* **18**, 11–22 (2010).
29. Glück, S. *et al.* *TP53* genomics predict higher clinical and pathologic tumor response in operable early-stage breast cancer treated with docetaxel-capecitabine ± trastuzumab. *Breast Cancer Res. Treat.* **132**, 781–791 (2012).
30. Curtis, C. *et al.* The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature* **486**, 346–352 (2012).
31. Cancer Genome Atlas Network. Comprehensive molecular portraits of human breast tumours. *Nature* **490**, 61–70 (2012).
32. Rhodes, D.R. *et al.* Oncomine 3.0: genes, pathways, and networks in a collection of 18,000 cancer gene expression profiles. *Neoplasia* **9**, 166–180 (2007).
33. Gray, K.A., Yates, B., Seal, R.L., Wright, M.W. & Bruford, E.A. Genenames.org: the HGNC resources in 2015. *Nucleic Acids Res.* doi:10.1093/nar/gku1071 (31 October 2014).
34. Chen, D. *et al.* LIFR is a breast cancer metastasis suppressor upstream of the Hippo-YAP pathway and a prognostic marker. *Nat. Med.* **18**, 1511–1517 (2012).
35. Gupta, R.A. *et al.* Long non-coding RNA *HOTAIR* reprograms chromatin state to promote cancer metastasis. *Nature* **464**, 1071–1076 (2010).
36. Prensner, J.R. *et al.* The long noncoding RNA *SChLAP1* promotes aggressive prostate cancer and antagonizes the SWI/SNF complex. *Nat. Genet.* **45**, 1392–1398 (2013).
37. Guttman, M. *et al.* Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. *Nature* **458**, 223–227 (2009).
38. Thomas, G. *et al.* A multistage genome-wide association study in breast cancer identifies two new risk alleles at 1p11.2 and 14q24.1 (*RAD51L1*). *Nat. Genet.* **41**, 579–584 (2009).
39. Stacey, S.N. *et al.* Common variants on chromosomes 2q35 and 16q12 confer susceptibility to estrogen receptor–positive breast cancer. *Nat. Genet.* **39**, 865–869 (2007).
40. Michailidou, K. *et al.* Large-scale genotyping identifies 41 new loci associated with breast cancer risk. *Nat. Genet.* **45**, 353–361 (2013).
41. Turnbull, C. *et al.* Genome-wide association study identifies five new breast cancer susceptibility loci. *Nat. Genet.* **42**, 504–507 (2010).
42. Li, J. *et al.* A combined analysis of genome-wide association studies in breast cancer. *Breast Cancer Res. Treat.* **126**, 717–727 (2011).
43. Amaral, P.P., Clark, M.B., Gascoigne, D.K., Dinger, M.E. & Mattick, J.S. lncRNAdb: a reference database for long noncoding RNAs. *Nucleic Acids Res.* **39**, D146–D151 (2011).
44. Volders, P.J. *et al.* LNCipedia: a database for annotated human lncRNA transcript sequences and structures. *Nucleic Acids Res.* **41**, D246–D251 (2013).
45. Park, C., Yu, N., Choi, I., Kim, W. & Lee, S. lncRNAtor: a comprehensive resource for functional investigation of long noncoding RNAs. *Bioinformatics* **30**, 2480–2485 (2014).
46. Hangauer, M.J., Vaughn, I.W. & McManus, M.T. Pervasive transcription of the human genome produces thousands of previously unidentified long intergenic noncoding RNAs. *PLoS Genet.* **9**, e1003569 (2013).
47. Zhou, Y. *et al.* Activation of p53 by *MEG3* non-coding RNA. *J. Biol. Chem.* **282**, 24731–24742 (2007).
48. Tomlins, S.A. *et al.* Urine *TMPRSS2:ERG* fusion transcript stratifies prostate cancer risk in men with elevated serum PSA. *Sci. Transl. Med.* **3**, 94ra72 (2011).
49. Prensner, J.R. *et al.* *PCAT-1*, a long noncoding RNA, regulates BRCA2 and controls homologous recombination in cancer. *Cancer Res.* **74**, 1651–1660 (2014).

ONLINE METHODS

High-performance computing. Computational analysis was performed using the Flux high-performance computer cluster hosted by Advanced Research Computing (ARC) at the University of Michigan.

RNA sequencing data processing. A comprehensive RNA-seq analysis pipeline was employed on all samples (**Supplementary Fig. 1b**). The analysis pipeline provided sequence quality metrics, filtering of contaminant reads, fragment size estimation, strand-specific library type estimation, spliced alignment of reads to the human reference genome (version hg19/GRCh37), alignment performance metrics, generation of visualization tracks for genome browsers and *ab initio* transcript assembly. The third-party tools used to process the RNA-seq data were selected on the basis of computational performance, ease of use, user and community support, and experience (**Supplementary Table 3**). Further details are described in the **Supplementary Note**.

Overview of transcriptome reconstruction. To merge the *ab initio*–assembled transcript fragments (transfrags) into a consensus transcriptome, we developed and used a bioinformatics method that (i) classifies and filters out sources of background noise in individual libraries and (ii) reassembles transfrags weighted by their expression levels from multiple libraries into a consensus transcriptome. More details are provided in the **Supplementary Note**.

Filtration of noise contamination. We controlled for alignment artifacts and poorly assembled transcripts by clipping very short first or last exons (<15 bp) and excluding short transfrags (≤250 bp). We removed noise due to genomic DNA contamination and incompletely processed RNA using a machine learning method. The method models the empirical distributions of relative transcript abundance and recurrence (number of independent samples in which the transcript was observed). From this model, the method determines the optimal library-specific thresholds for distinguishing annotated from unannotated transcription as a proxy for signal versus background noise, respectively. Further details are described in the **Supplementary Note**.

Transcriptome meta-assembly. We created directed acyclic splicing graphs where nodes in the graph reflected contiguous exonic regions and edges corresponded to splicing possibilities (**Supplementary Fig. 3a**). Nodes in the splicing graph with relatively low abundance were then pruned. We then incorporated the partial path information inherent for transfrags spanning multiple exons by building splicing pattern graphs that subsumed the original splice graphs (**Supplementary Fig. 3b**). The splicing pattern graph is a type of de Bruijn graph where each node represents a contiguous path of length k through the splice graph and edges connect paths with $k - 1$ nodes in common. The algorithm finds and reports a set of highly abundant transcripts by iteratively traversing the graph using dynamic programming in a greedy fashion. Further details are described in the **Supplementary Note**.

Merging of meta-assemblies. To merge the meta-assemblies from 18 cohorts, we used the Cuffmerge tool[50], which produced a final transcriptome GTF file.

Comparisons of MiTranscriptome with reference catalogs. The exons, splice sites and splicing patterns of all assembled transcripts were compared to RefSeq, UCSC, GENCODE (version 19) and the merged union of all three reference catalogs using custom Python scripts. Sensitivity and precision values were computed using the number of shared strand-specific transcribed bases, introns and splicing patterns. Precision was also computed for the subset of *ab initio* transcripts that overlapped any part of a reference transcript. Transcripts that overlapped a reference transcript on the same strand were designated as annotated. When an *ab initio* transcript matched multiple reference transcripts, a best match was chosen using the following criteria: (i) matching splicing patterns, (ii) fraction of shared introns and (iii) fraction of shared transcribed bases. The biotype (protein, read-through, pseudogene or lncRNA) for the annotated transcripts was imputed from the best match reference transcript. Annotated lncRNAs and unannotated transcripts were reclassified as either lncRNAs or TUCPs.

Prediction of transcripts of unknown coding potential. We predicted coding potential by integrating two sources of evidence: (i) predictions from the alignment-free Coding Potential Assessment Tool (CPAT)[14] and (ii) searches for Pfam 27.0 matches[15]. CPAT determines the coding probability of transcript sequences using a logistic regression model built from ORF size, Fickett TESTCODE statistic[51] and hexamer usage bias. We chose a CPAT probability cutoff by repeatedly randomly sampling 100,000 each of putative noncoding and protein-coding transcripts and optimizing on the balanced accuracy metric (average of sensitivity and specificity; **Supplementary Fig. 8b,c**). The average area under the curve (AUC) across 100 iterations was 0.9310 (minimum = 0.9302, maximum = 0.9320), and the average optimal probability cutoff was 0.5242 (minimum = 0.5090, maximum = 0.5482). This cutoff value achieved accurate discrimination of lncRNAs and protein-coding genes (sensitivity = 0.84, specificity = 0.95, FDR = 0.076). Of the putative noncoding transcripts, 9,903 (5.3%) exceeded the CPAT cutoff and met the criteria for TUCPs. As additional evidence of coding potential, we scanned all transcripts for Pfam A or B domains across the three translated reading frames for stranded transcripts and the six reading frames for monoexonic transcripts of unknown strand (**Supplementary Note**). We designated putative noncoding transcripts with either a Pfam domain or a positive CPAT prediction as TUCPs.

Proteomics analysis. We obtained the following Thermo files (in the RAW format) from a recent study mapping the human proteome[52]: Adult_Kidney_Gel_Elite_55, Adult_Liver_Gel_Elite_56, Adult_Pancreas_Gel_Elite_60, Adult_Rectum_Gel_Elite_63, Adult_Urinarybladder_Gel_Elite_40, Fetal_Brain_Gel_Velos_16, Adult_Lung_Gel_Elite_56 and Adult_Prostate_Gel_Elite_62. The Thermo files were transformed into mzXML using MSConverter[53] and interrogated against human UniProt database V.15.11 using the X!Tandem search engine. The database was concatenated with all possible ORFs longer than 7 amino acids from the lncRNAs and with reversed sequences for the determination of FDR values. The X!Tandem search parameters were as follows: fully tryptic cleavage, parent mass error 5 ppm, fragment mass error 0.5 Da, 2 allowed missed cleavages, fixed modifications: cysteine carbamidomethylation; variable modifications: methionine oxidation. X!Tandem output files were processed by PeptideProphet and ProteinProphet. Data were filtered at a peptide probability of 0.5 and a protein probability of 0.9 to ensure protein FDR < 1%.

Confidence scoring system. After assembly of the MiTranscriptome, transcripts were subjected to an additional confidence evaluation. lncRNAs in the MiTranscriptome were categorized into tiers on the basis of their annotation status and the degree of matching of splice junctions to the reference annotation (**Supplementary Table 4**). Tier 1 transcripts are all annotated, and tier 2 transcripts are unannotated. An empirical cumulative distribution function (eCDF) was developed by profiling the second highest expression value (across all 6,503 samples) for each tier 1 transcript. The second highest value was used to control for outlier expression. The eCDF was used to compute confidence scores for tier 2 transcripts using the same expression summary statistic.

Validation of lncRNA transcripts by quantitative RT-PCR. We chose 150 lncRNAs with at least 1 FPKM expression in either A549, LNCaP or MCF-7 cells for biological validation. For each transcript, primer pairs were designed using the Primer-BLAST tool[54]. Primer pairs with the following parameters were selected: (i) amplicon length of 80–140 bp, (ii) primer GC content of 35–65% and (iii) primer length greater than 20 bp. Primers were used for BLAST runs against the human genome to ensure specificity to our target gene, and primers designed against multiexonic transcripts spanned exon junctions. Regions of any transcript that directly overlapped an exon on the antisense strand were avoided. Primer pairs meeting these criteria could be designed for 100 of 150 lncRNAs (38 monoexonic and 62 multiexonic). All oligonucleotide primers were obtained from Integrated DNA Technologies, and their sequences are listed in **Supplementary Table 5**.

RNA was isolated from A549, LNCaP and MCF-7 cells in TRIzol (Invitrogen) using the RNeasy Mini kit (Qiagen). An equal amount of RNA was converted to cDNA using random primers and the SuperScript III reverse transcription system (Invitrogen). qRT-PCR was performed using Power SYBR Green Mastermix (Applied Biosystems) on an Applied Biosystems 7900HT Real-Time PCR System. The housekeeping genes *CHMP2A*, *EMC7*, *GPI*, *PSMB2*, *PSMB4*, *RAB7A*, *REEP5* and *SNRPD3* were used as loading controls[55]. Data were normalized first to the values for housekeeping genes and then to the median value for all samples using the $\Delta\Delta C_t$ method and plotted as fold change over the median. To ensure the specificity of the primers, 20 amplicons were further analyzed by Sanger sequencing.

Cell lines and reagents. All cell lines were obtained from the American Type Culture Collection (ATCC). Cell lines were maintained using standard conditions. Specifically, A549 cells were grown in F-12K with 10% FBS, LNCaP cells were grown in RMPI-1640 (Invitrogen) with 10% FBS and 1% penicillin-streptomycin and MCF-7 cells were grown in Eagle's Minimum Essential Medium (EMEM) plus 10% FBS. All of the cell lines were grown in a cell culture incubator at 37 °C with 5% CO_2. To ensure their identity, cell lines were genotyped at the University of Michigan Sequencing Core using Profiler Plus (Applied Biosystems) and compared with the short tandem repeat (STR) profiles of respective cell lines available in the STR Profile Database (ATCC). All of the cell lines were routinely tested and found to be free of Mycoplasma contamination.

Evidence for active regulation of transcriptional start sites. To conduct analysis of TSS intervals, ENCODE Project data sets were downloaded from the UCSC Genome Browser[13]. For H3K4me3 analysis, we used the ENCODE Project Broad Institute H3K4me3 ChIP-seq peaks for the GM12878, H1-hESC, HeLa-S3, HepG2, HMEC, HSMM, HSMMtube, HUVEC, K562, NH-A, NHDF-Ad, NHEK and NHLF cell lines[56]. For Pol II analysis we used POL2RA binding sites from the ENCODE Project Uniform TFBS master file version 3 for any of the cell lines with H3k4me3 data[19]. Finally, for the DNase I hypersensitivity analysis, the ENCODE Project combined University of Washington and Duke University DNase I hypersensitivity regions were downloaded as a master file from the European Molecular Biology Laboratory–European Bioinformatics Institute (EMBL-EBI) and filtered for any of the cell lines with H3k4me3 data. Peak enrichment files (BED format) were aggregated across all cell lines. Intervals of ±10 kb surrounding unique MiTranscriptome TSSs were generated using the BEDTools slop tool[57]. To control for expression, TSSs were filtered to remove transcripts not expressed in any of the cell lines (FPKM < 0.1). Base-wise peak coverage was generated for each TSS interval using the BEDTools coverage function and summarized across subsets of TSSs. Summed per-base coverage histograms were normalized by dividing by the number of expressed TSSs.

Conservation analysis. The evolutionary conservation of transcripts in our assembly was studied using two metrics: (i) the fraction of significantly conserved bases ($P \leq 0.01$, phyloP algorithm) and (ii) the maximally conserved 200-nt sliding window (phastCons scores averaged within each window). The former captures independently conserved elements within a transcript regardless of position, and the latter captures contiguous regions of high conservation. The 200-nt sliding window size was chosen to aid in the discovery of putative UCEs[23]. As a negative control, we measured the conservation of non-transcribed regions using these metrics by randomly sampling contiguous length-matched intervals from intergenic and intronic space. Non-transcribed interval sampling was restricted to regions with valid 46-way conservation data.

The fractional base-wise conservation and contiguous window conservation metrics were used to nominate highly conserved and ultraconserved transcripts, respectively. In both cases, cutoffs for significant transcripts were determined by controlling the rate of observing elements with similar conservation levels within non-transcribed intergenic space at a level of 0.01. For fractional base-wise conservation, a score of 0.0947 (9.5% of transcript bases conserved at phyloP P value < 0.01) corresponded to FDR < 0.01. At this cutoff, the sensitivity for detecting protein-coding transcripts was 0.67. For contiguous sliding window conservation, an average phastCons probability of 0.9986 corresponded to FDR < 0.01. At this cutoff, the sensitivity for detecting true positive ultraconserved noncoding elements downloaded from UCNEbase was 0.69 (ref. 22). Applying these criteria to our assembly yielded 6,034 lncRNAs (3.4%) and 541 TUCPs (4.7%) with significant base-wise conservation levels. Additionally, 1,686 lncRNAs (0.96%) and 121 TUCPs (0.01%) harbored contiguous ultraconserved regions.

GWAS analysis. A list of GWAS SNPs was obtained from the National Human Genome Research Institute's GWAS catalog (accessed 6 January 2014)[24]. SNP haplotypes were excluded from the SNP overlap analysis, and a list of 11,194 unique SNPs was obtained. The merged union of the RefSeq, UCSC and GENCODE catalogs was used as a reference for comparison with MiTranscriptome. Please refer to the **Supplementary Note** for a description of the GWAS overlap enrichment testing analysis.

Transcript expression estimation. Expression levels (FPKM) of the transcripts in the assembly were determined using Cufflinks (versions 2.02 and 2.1.1)[58]. Normalized abundance estimates (FPKM) were computed for all MiTranscriptome transcripts, converted into approximate fragment count values and aggregated into a matrix of expression data (**Fig. 4a** and **Supplementary Note**). Library size factors for expression normalization were computed by applying the geometric normalization method described by Anders and Huber[59].

Transcript expression enrichment analysis. To analyze the differential expression of transcripts relative to sample phenotypes, we developed a method called Sample Set Enrichment Analysis (SSEA). SSEA performs weighted KS tests using normalized count data vectors as weights. To convert count values into weights for a single KS test, the following steps are performed: (i) raw count values are normalized by library-specific size factors, (ii) normalized count values are 'resampled' from a Poisson distribution (where λ equals the observed count value) to mimic the effect of technical replication and (iii) random Poisson noise (by default, $\lambda = 1$) is added to the normalized, resampled count values to destabilize zero-valued counts and break ties. A power transform (exponential or logarithmic) is then applied to the weights (by default, a log transformation is applied after incrementing normalized count values by 1). The choice of power transformation influences the relative importance of precision versus recall during enrichment testing. For example, users aiming to discover genes new in molecular subtypes of a disease would prioritize precision over sensitivity, whereas a user aiming to discover ideal biomarkers might value sensitivity over precision. After count data normalization and power transformation, SSEA performs the weighted KS test procedure described in GSEA[25,60]. The resulting enrichment score statistic describes the strength of the association between the weights and the sample set.

To control for random sampling bias in count values (for example, 'shot noise'), SSEA performs repeated enrichment tests using resampled count values to mimic observations from technical replicates and uses the median enrichment score (by default, 100 tests are performed). The basis for Poisson resampling as a legitimate model for technical replication was established by Marioni *et al.*[60]. To test for significance, SSEA performs enrichment tests using randomly shuffled sample labels to derive a set of null enrichment scores with the same sign as the observed score (by default, 1,000 null enrichment scores are computed). The nominal P value reported is the relative rank of the observed enrichment score within the null enrichment scores. To control for multiple-hypothesis testing, SSEA maintains the null normalized enrichment score (NES) distributions for all transcripts in a sample set and uses the null NES distribution to compute FDR q values in the same manner as proposed by Subramanian *et al.*[25].

Benchmarking SSEA performance using microarray gene signatures. Gene signatures for the top 1% of overexpressed and underexpressed genes from three prostate cancer[26–28] and three breast cancer[29–31] microarray studies were obtained using Oncomine[32] (**Supplementary Table 8**). The top 1% of gene signatures as detected by SSEA in the MiTranscriptome breast and prostate cohorts were determined using prostate cancer versus normal and breast cancer versus normal sample sets (**Fig. 4a**). Given that the MiTranscriptome was produced from an *ab initio* assembly, transcript identity was assigned to the annotated reference gene with the greatest degree of concordance, where degree of splicing agreement was prioritized

doi:10.1038/ng.3192

over degree of exonic same-stranded overlap. The most enriched isoform for each gene was used to produce a gene signature.

The degree of overlap for all combinations of the 16 gene sets tested (3 published breast upregulated sets, 3 published breast downregulated sets, 3 published prostate upregulated sets, 3 published prostate downregulated sets, 1 SSEA-determined prostate upregulated set, 1 SSEA-determined prostate downregulated set, 1 SSEA-determined breast upregulated set and 1 SSEA-determined breast downregulated set) was determined by calculating an odds ratio and performing a Fisher's exact test for each gene set pair (**Supplementary Table 9**). Each comparison was restricted to the set of the genes assessed by both profiling platforms. Microarray chip annotation files were downloaded from the Molecular Signatures Database (MSigDB)[61]. The set of all annotated genes (relative to RefSeq, UCSC and GENCODE) was used as the annotation file for MiTranscriptome. Unsupervised hierarchical clustering of the heat map data was performed using the 'euclidean' distance measure and the 'complete' agglomeration method.

Discovery of lineage-specific and cancer-specific transcripts. To generate enrichment test data for unsupervised clustering, we ranked transcripts within each SSEA sample set by NES and assigned fractional ranks (for example, a fractional rank of 0.95 implies the transcript ranked in the top 5th percentile of all transcripts in the sample set). Only significant results (FDR $< 1 \times 10^{-7}$ for lineage analysis and FDR $< 1 \times 10^{-3}$ for cancer versus normal analysis) were used. Unsupervised clustering was performed using Pearson correlation of log-transformed fractional ranks as a distance metric and Ward's method. Transcripts that were significantly associated with multiple sample sets were grouped with the most strongly associated sample set. Heat maps were produced using the heatmap.2 function from the gplots package in R.

Guilt-by-association GSEA analysis. For each cancer- and/or lineage-associated lncRNA (**Supplementary Table 10**), expression levels of the target lncRNA were correlated to the expression of all protein-coding genes across all samples in the associated tissue cohort. For cancer cohorts (for example, breast and prostate cancers), correlations were performed (Spearman) using only the cancer samples (normal samples were excluded). Protein-coding genes were then ranked by ρ value and used in a weighted, pre-ranked GSEA analysis against a collection of cancer-associated gene sets from MSigDB (**Supplementary Table 11**). Significant associations were determined for any gene set having an FWER P value below 0.001.

50. Trapnell, C. *et al.* Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat. Protoc.* **7**, 562–578 (2012).
51. Fickett, J.W. Recognition of protein coding regions in DNA sequences. *Nucleic Acids Res.* **10**, 5303–5318 (1982).
52. Kim, M.S. *et al.* A draft map of the human proteome. *Nature* **509**, 575–581 (2014).
53. Chambers, M.C. *et al.* A cross-platform toolkit for mass spectrometry and proteomics. *Nat. Biotechnol.* **30**, 918–920 (2012).
54. Ye, J. *et al.* Primer-BLAST: a tool to design target-specific primers for polymerase chain reaction. *BMC Bioinformatics* **13**, 134 (2012).
55. Eisenberg, E. & Levanon, E.Y. Human housekeeping genes, revisited. *Trends Genet.* **29**, 569–574 (2013).
56. Bernstein, B.E. *et al.* Genomic maps and comparative analysis of histone modifications in human and mouse. *Cell* **120**, 169–181 (2005).
57. Quinlan, A.R. & Hall, I.M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).
58. Trapnell, C. *et al.* Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat. Biotechnol.* **28**, 511–515 (2010).
59. Anders, S. & Huber, W. Differential expression analysis for sequence count data. *Genome Biol.* **11**, R106 (2010).
60. Marioni, J.C., Mason, C.E., Mane, S.M., Stephens, M. & Gilad, Y. RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. *Genome Res.* **18**, 1509–1517 (2008).
61. Liberzon, A. *et al.* Molecular signatures database (MSigDB) 3.0. *Bioinformatics* **27**, 1739–1740 (2011).

doi:10.1038/ng.3192

The lncRNAs PCGEM1 and PRNCR1 are not implicated in castration resistant prostate cancer

John R. Prensner[1,*], Anirban Sahu[1,*], Matthew K. Iyer[1,2,*], Rohit Malik[1,*], Benjamin Chandler[1], Irfan A. Asangani[1], Anton Poliakov[1], Ismael A. Vergara[3], Mohammed Alshalalfa[3], Robert B. Jenkins[4], Elai Davicioni[3], Felix Y. Feng[1,5,7], Arul M. Chinnaiyan[1,2,6,7,8]

[1] Michigan Center for Translational Pathology, University of Michigan, Ann Arbor, Michigan USA.

[2] Department of Computational Medicine and Bioinformatics, Ann Arbor, Michigan USA.

[3] GenomeDx Biosciences Inc., Vancouver, British Columbia, Canada.

[4] Department of Laboratory Medicine and Pathology, Mayo Clinic, Rochester, Minnesota USA.

[5] Department of Radiation Oncology, University of Michigan, Ann Arbor, Michigan USA.

[6] Department of Pathology, University of Michigan, Ann Arbor, Michigan USA.

[7] Comprehensive Cancer Center, University of Michigan, Ann Arbor, Michigan USA.

[8] Howard Hughes Medical Institute, University of Michigan, Ann Arbor, Michigan USA.

[*] These authors contributed equally

Correspondence to: *Arul M. Chinnaiyan,* ***email:*** *arul@med.umich.edu*

Keywords: prostate cancer, long noncoding RNA, androgen receptor

Received: *February 5, 2014* ***Accepted***: *March 21, 2014* ***Published***: *March 23, 2014*

ABSTRACT:

Long noncoding RNAs (lncRNAs) are increasingly implicated in cancer biology, contributing to essential cancer cell functions such as proliferation, invasion, and metastasis. In prostate cancer, several lncRNAs have been nominated as critical actors in disease pathogenesis. Among these, expression of *PCGEM1* and *PRNCR1* has been identified as a possible component in disease progression through the coordination of androgen receptor (AR) signaling (Yang et al., Nature 2013, see ref. [1]). However, concerns regarding the robustness of these findings have been suggested. Here, we sought to evaluate whether *PCGEM1* and *PRNCR1* are associated with prostate cancer. Through a comprehensive analysis of RNA-sequencing data (RNA-seq), we find evidence that *PCGEM1* but not *PRNCR1* is associated with prostate cancer. We employ a large cohort of >230 high-risk prostate cancer patients with long-term outcomes data to show that, in contrast to prior reports, neither gene is associated with poor patient outcomes. We further observe no evidence that *PCGEM1* nor *PRNCR1* interact with AR, and neither gene is a component of AR signaling. Thus, we conclusively demonstrate that *PCGEM1* and *PRNCR1* are not prognostic lncRNAs in prostate cancer and we refute suggestions that these lncRNAs interact in AR signaling.

INTRODUCTION

Long noncoding RNAs (lncRNAs) have emerged as a critical element in cell biology, contributing to a wide variety of cellular behaviors and functions [2]. In cancer, lncRNAs have been the subject of much research during the past five years. Notably, lncRNAs are known to coordinate aggressive phenotypes of several common tumors, including breast cancer and prostate cancer [3, 4]. Large profiling studies have suggested that upwards of 10,000 lncRNAs may exist in the human genome [5]; yet only a fraction of these entities have been characterized. Thus, the identity and function of lncRNAs in cancer remains largely unknown.

In prostate cancer, several lncRNAs, including *PCA3* and *PCAT-1*, have been shown to be upregulated in patients with cancer [6-9]. Recently, two lncRNAs, *PCGEM1* and *PRNCR1*, have been suggested in prostate cancer to act as mediators of castration-resistance disease by binding, in a direct and sequential fashion, to the androgen receptor (AR), causing ligand-independent activation of its gene expression programs [1]. While *PCGEM1* has been observed in prostate cancer previously [6, 10], *PRNCR1* is a poorly characterized transcript, and we were concerned that *PRNCR1* had not been nominated by previous global profiling studies of prostate cancers [7, 11-14].

We therefore sought to investigate *PRNCR1* and *PCGEM1* in prostate cancer. In specific, we sought to reproduce three core observations suggested by Yang et al published in *Nature* (see [1]) and include: 1) that *PRNCR1* and *PCGEM1* are highly overexpressed in aggressive forms of prostate cancer, 2) that these two lncRNAs bind to AR under ligand-stimulated conditions, and 3) that the coordination of *PRNCR1* and *PCGEM1* interact with AR via specific post-translational modifications of the AR protein. Here, we report that none of these three findings is fully reproducible.

First, we asked whether *PCGEM1* and *PRNCR1* are highly overexpressed in aggressive prostate cancer, as suggested by others (see [1, 15]). Indeed, while some have argued that these lncRNAs are critical in castration-resistant prostate cancer [1], there has been no study that evaluated the expression of these lncRNAs in tissue samples from human castrate-resistant prostate cancers (CRPC). To evaluate these lncRNAs in more detail, we first assessed their expression levels in 171 human prostatic tissues using RNA sequencing data aggregated from four independent studies of prostate cancer, including our own internal datasets [1, 12-14] (Fig. 1A). Whereas we found robust expression of *PCGEM1* in a subset of prostate tissues (RPKM >1 in 82 samples; RPKM >10 in 27 samples), we observed scant levels of *PRNCR1* in all samples (RPKM >1 in only 3 samples; RPKM >10 in 0 samples) (Supplementary Table 1). This does not lend confidence to *PRNCR1* as a significant entity in this disease. For comparison, we used the prostate cancer lncRNA *SChLAP1* as a positive control. We found extreme overexpression of *SChLAP1* in samples from all datasets (RPKM >1 in 69 samples; RPKM >10 in 26 samples) (Supplementary Fig. 1). To rule out the possibility that *PRNCR1* was a non-poly-adenylated RNA, we verified experimentally that *PRNCR1* was observed in the poly-A fraction of RNA that was used to generate the RNA-seq data (Supplementary Fig. 2).

Given the low support for *PRNCR1* in the RNA-seq data, we next confirmed these findings using quantitative PCR (qPCR) in a large set of prostate cancer tissues including 34 PCAs and 31 CRPC tumors as well as 18 benign adjacent tissues. As shown in Fig. 1B, *PCGEM1* is upregulated in clinically localized cancer, confirming the known literature [6, 10]; however *PRNCR1* expression does not demonstrate a convincing association with prostate cancer. We found a borderline decrease in *PRNCR1* expression in metastatic castrate-resistant cancer (p = 0.047, Student's t-test). We used *PCAT1*, *EZH2*, and *SChLAP1* as control genes, all of which have elevated expression in prostate cancer metastases. Conversely, we used *PCA3* as a control gene that is known to be upregulated in localized prostate cancer but not metastatic prostate cancer. Finally, while *PCGEM1* is upregulated in cancer patients from matched tumor/benign samples, *PRNCR1* does not convincingly exhibit this pattern of upregulation (Supplementary Fig. 3).

Next, an independent analysis of 235 high-risk prostate cancer tissues demonstrated that neither *PCGEM1* nor *PRNCR1* is associated with aggressive prostate cancer, and neither lncRNA stratifies prostate cancer-specific mortality (Fig. 1C,D and Supplementary Tables 2,3). An analysis of intermediate endpoints such as biochemical recurrence and progression to metastatic disease demonstrated a trend for *PCGEM1* and *PRNCR1* to be associated with less aggressive disease and favorable outcomes (Supplementary Fig. 4), which contradicts previous claims that these lncRNAs are involved in an aggressive patient clinical course [1, 15]. Using an independent validation cohort of tissues we verified that neither *PCGEM1* nor *PRNCR1* is associated with aggressive prostate cancer (Supplementary Table 2). By contrast, we have used these datasets to confirm the prognostic utility of the lncRNA *SChLAP1* in prostate cancer, and high expression of *SChLAP1* is a powerful predictor for poor patient survival (Fig. 1E) [4].

Next, we examined whether *PCGEM1* and *PRNCR1* interacted with AR. We performed RNA-IP (RIP) assays using two independent AR antibodies, including the same antibody that was previously used to show an interaction between these lncRNAs and AR [1]. In accordance with the published literature, we performed a time-series of RIP experiments following AR stimulation, because prior data suggests that these lncRNAs bind AR from 1-2 hours after AR stimulation but not at 4 hours post-stimulation [1]. In our RIP experiments, we could not confirm that AR binds to *PCGEM1* or *PRNCR1* at either 1 hour or 4 hours post-stimulation with DHT (Fig. 1F and Supplementary Fig. 5). Similarly, in cells grown at steady-state, we used a second AR antibody and did not observe binding between AR and *PCGEM1* or *PRNCR1* (Supplementary Fig. 6). DHT-stimulated cells also demonstrated no induction in *PCGEM1* or *PRNCR1* expression (Supplementary Fig. 7). These results imply that *PCGEM1* and *PRNCR1* are not AR-interacting lncRNAs.

Finally, earlier data propose that *PCGEM1* and *PRNCR1* interact with AR via specific post-translational modifications (PTMs), specifically K349 methylation (K349Me) for *PCGEM1* and K631/K634 acetylation (K631Ac/K634Ac) for *PRNCR1* [1]. To search for these PTMs, we independently performed mass spectrometry for

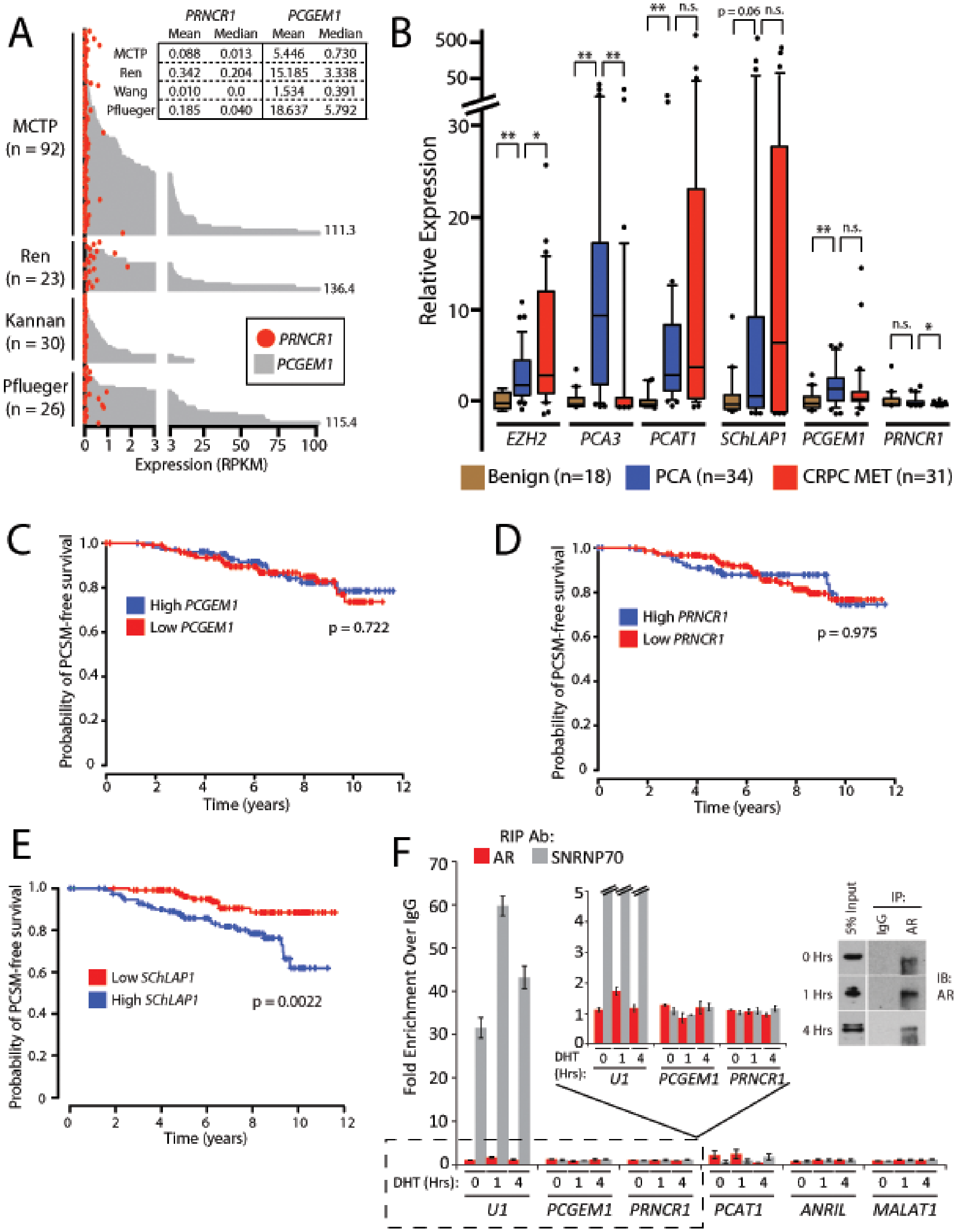


Figure 1: PCGEM1 and PRNCR1 are not associated with prostate cancer progression and do not bind the androgen receptor. (A) Plot showing *PCGEM1* (grey bars) and *PRNCR1* (red circles) expression levels (Reads per Kilobase per Million Reads, or RPKM) across 171 samples from four RNA-Seq studies of prostate cancer: Michigan Center for Translational Pathology (MCTP, internal data and dbGAP, phs000443.v1.p1), Ren et al. [13] (EGA, ERP00550), Kannan et al. [14] (GEO, GSE22260), and Pflueger et al. [12] (dbGAP, phs000310.v1.p1). Inset box shows descriptive statistics for each study. (B) Quantitative PCR for *PCGEM1* and *PRNCR1* in a cohort of prostate cancer tissues, benign (n = 18), localized cancer (n =34), metastatic cancer (n = 31). An asterisk (*) indicates $p < 0.05$. Two asterisks (**) indicate $p < 0.01$. n.s. = non-significant. P values were determined by a two-tailed Student's t-test. Data for *SChLAP1* is obtained and re-analyzed from a prior publication (ref. [4]). (C) *PCGEM1* expression does not predict for prostate cancer-specific mortality (PCSM). (D) *PRNCR1* expression does not predict for PCSM. (E) High *SChLAP1* expression is a powerful predictor of PCSM (p = 0.0022). Data in (E) is reproduced from a prior publication (ref. [4]). P values in (C-E) are determined using a log-rank test. (F) RNA-immunoprecipitation (RIP) for AR following stimulation of LNCaP cells with 100nM DHT does not show binding of *PRNCR1* or *PCGEM1* to AR. *U1* binding to SNRNP70 is used as a positive control. *PCAT-1*, *ANRIL*, and *MALAT1* serve as negative controls. Inset: Western blot confirmation of AR protein pull-down by the immunoprecipitation assays. Error bars represent S.E.M.

AR in the LNCaP cell line, achieving 95% coverage of all possible tryptic peptides. We were unable to confirm that these PTMs (K349Me, K631Ac, or K634Ac) are present on AR (Supplementary Fig. 8 and Supplementary Table 4). To examine this discrepancy further, we re-analyzed prior AR MS data (found in [1]). Although this MS dataset was obtained with a trypsin digestion to prepare samples for MS, we found no fully tryptic peptides supporting the nomination of K349Me, K631Ac, or K634Ac (Supplementary Fig. 8). In fact, in the MS data for these PTMs in ref. [1], almost all peptides harboring these PTMs are non-tryptic, which are generally considered to be analysis artifacts since true non-tryptic peptides are exceedingly rare following a trypsin digestion [16-18] (Supplementary Discussion). Non-tryptic peptides are also associated with a high false-discovery rate [19]. All peptides nominating the K349Me, K631Ac, or K634Ac PTMs in ref. [1] also had multiple additional PTMs that were nominated, indicating non-specificity. These included extraordinarily rare and unusual PTMs such as oxidated lysine and deamidated asparagine, which suggest technical artifacts given the negligible likelihood of multiple rare and unusual PTMs occurring on true non-trypic peptides. The statistical confidence for these non-tryptic peptides is <5%, whereas the corresponding fully tryptic peptides for these amino acid residues had statistical confidences >90%.

In summary, we have been unable to show a convincing role for *PCGEM1* or *PRNCR1* in aggressive prostate cancer or AR signaling. First, our data analysis of numerous human prostate cancer tissues from multiple independent laboratories indicates that neither *PCGEM1* nor *PRNCR1* are associated with castration-resistant prostate cancer. Second, we were unable to verify that *PCGEM1* and *PRNCR1* bind to the androgen receptor. Lastly, we are unconvinced that the K349Me, K631Ac, or K634Ac AR PTMs represent a plausible mechanism for interaction between AR and *PCGEM1* and *PRNCR1*. While our results challenge the notion that *PCGEM1* and *PRNCR1* play a causal role in prostate cancer, we regard lncRNAs as an emerging field of study in cancer [3, 6, 20, 21] and we are encouraged by the interest in lncRNAs in prostate cancer.

METHODS

Prostate tissues were obtained from the radical prostatectomy series and Rapid Autopsy Program at the University of Michigan tissue core. All tissue samples were collected with informed consent under an Institutional Review Board (IRB) approved protocol at the University of Michigan. Outcomes analyses were performed on a cohort of Mayo Clinic prostate cancer radical prostatectomy samples obtained under an IRB-approved protocol as described previously. Cell lines were maintained according to standard conditions. For RIP experiments, cells were deprived of androgen for 48 hours prior to stimulation with 100nM DHT. RIP experiments were performed as previously described [1, 4]. Bioinformatics analyses utilized publicly available RNA-Seq data. Please see Supplementary Methods for details.

ACKNOWLEDGEMENTS

We thank Xia Jiang and Shruthi Subramaniam for technical assistance. This work was supported in part by the NIH Prostate Specialized Program of Research Excellence grant P50CA69568, the Early Detection Research Network grant UO1 CA111275, the US National Institutes of Health R01CA132874-01A1, and the Department of Defense grant PC100171 (A.M.C.). A.M.C. is supported by the Doris Duke Charitable Foundation Clinical Scientist Award, the Prostate Cancer Foundation, and the Howard Hughes Medical Institute. A.M.C. is an American Cancer Society Research Professor and a Taubman Scholar of the University of Michigan. F.Y.F. was supported by the Prostate Cancer Foundation, the Department of Defense grant PC094231. J.R.P. was supported by a Prostate Cancer Foundation Young Investigator Award. A.S. was supported by the NIH Predoctoral Fellowship 1F30CA180376-01. M.K.I. was supported by the Department of Defense Predoctoral Fellowship BC100238. R.M. was supported by the Department of Defense Post-doctoral Fellowship W81XWH-13-1-0284. J.R.P., M.K.I., and A.S. are Fellows of the University of Michigan Medical Scientist Training Program.

Disclosures and Competing Financial Interests

The University of Michigan has filed a patent on lncRNAs in prostate cancer, including *SChLAP1*, in which A.M.C.,J.R.P. and M.K.I. are named as co-inventors. Wafergen, Inc. has a non-exclusive license for creating commercial research assays for lncRNAs in prostate cancer. GenomeDx Biosciences Inc has an exclusive license for creating tissue assays for lncRNAs in prostate cancer. A.M.C. is a co-founder and advisor to Compendia Biosciences, which supports the Oncomine database. He also serves on the Scientific Advisory Board of Wafergen; neither Life Technologies or Wafergen had any role in the design or experimentation of this study, nor have they participated in the writing of the manuscript. I.A.V. and E.D. are employees of GenomeDx Biosciences Inc.

Author Contributions

J.R.P., R.M., M.K.I., A.S. and A.M.C. designed the project and directed experimental studies. J.R.P, R.M, A.S. and B.C. performed *in vitro* studies. M.K.I. performed bioinformatics analysis. I.A.A. and A.P. performed AR mass spectrometry. I.A.V., R.B.J., E.D., and M.A. acquired human tissue samples and performed statistical outcomes analyses for *PCGEM1* and *PRNCR1* expression. J.R.P., M.K.I., A.S., R.M., F.Y.F. and A.M.C. designed experiments, interpreted data, and wrote the manuscript.

REFERENCES

1. Yang L, Lin C, Jin C, Yang JC, Tanasa B, Li W, Merkurjev D, Ohgi KA, Meng D, Zhang J, Evans CP and Rosenfeld MG. lncRNA-dependent mechanisms of androgen-receptor-regulated gene activation programs. Nature. 2013; 500(7464):598-602.
2. Rinn JL and Chang HY. Genome regulation by long noncoding RNAs. Annu Rev Biochem. 2012; 81:145-166.
3. Gupta RA, Shah N, Wang KC, Kim J, Horlings HM, Wong DJ, Tsai MC, Hung T, Argani P, Rinn JL, Wang Y, Brzoska P, Kong B, Li R, West RB, van de Vijver MJ, et al. Long non-coding RNA HOTAIR reprograms chromatin state to promote cancer metastasis. Nature. 2010; 464(7291):1071-1076.
4. Prensner JR, Iyer MK, Sahu A, Asangani IA, Cao Q, Patel L, Vergara IA, Davicioni E, Erho N, Ghadessi M, Jenkins RB, Triche TJ, Malik R, Bedenis R, McGregor N, Ma T, et al. The long noncoding RNA SChLAP1 promotes aggressive prostate cancer and antagonizes the SWI/SNF complex. Nat Genet. 2013; 45(11):1392-1398.
5. Cabili MN, Trapnell C, Goff L, Koziol M, Tazon-Vega B, Regev A and Rinn JL. Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. Genes Dev. 2011; 25(18):1915-1927.
6. Du Z, Fei T, Verhaak RG, Su Z, Zhang Y, Brown M, Chen Y and Liu XS. Integrative genomic analyses reveal clinically relevant long noncoding RNAs in human cancer. Nat Struct Mol Biol. 2013; 20(7):908-913.
7. Prensner JR, Iyer MK, Balbin OA, Dhanasekaran SM, Cao Q, Brenner JC, Laxman B, Asangani IA, Grasso CS, Kominsky HD, Cao X, Jing X, Wang X, Siddiqui J, Wei JT, Robinson D, et al. Transcriptome sequencing across a prostate cancer cohort identifies PCAT-1, an unannotated lincRNA implicated in disease progression. Nat Biotechnol. 2011; 29(8):742-749.
8. de Kok JB, Verhaegh GW, Roelofs RW, Hessels D, Kiemeney LA, Aalders TW, Swinkels DW and Schalken JA. DD3(PCA3), a very sensitive and specific marker to detect prostate tumors. Cancer Res. 2002; 62(9):2695-2698.
9. Hessels D and Schalken JA. The use of PCA3 in the diagnosis of prostate cancer. Nat Rev Urol. 2009; 6(5):255-261.
10. Srikantan V, Zou Z, Petrovics G, Xu L, Augustus M, Davis L, Livezey JR, Connell T, Sesterhenn IA, Yoshino K, Buzard GS, Mostofi FK, McLeod DG, Moul JW and Srivastava S. PCGEM1, a prostate-specific gene, is overexpressed in prostate cancer. Proc Natl Acad Sci U S A. 2000; 97(22):12216-12221.
11. Taylor BS, Schultz N, Hieronymus H, Gopalan A, Xiao Y, Carver BS, Arora VK, Kaushik P, Cerami E, Reva B, Antipin Y, Mitsiades N, Landers T, Dolgalev I, Major JE, Wilson M, et al. Integrative genomic profiling of human prostate cancer. Cancer Cell. 2010; 18(1):11-22.
12. Pflueger D, Terry S, Sboner A, Habegger L, Esgueva R, Lin PC, Svensson MA, Kitabayashi N, Moss BJ, MacDonald TY, Cao X, Barrette T, Tewari AK, Chee MS, Chinnaiyan AM, Rickman DS, et al. Discovery of non-ETS gene fusions in human prostate cancer using next-generation RNA sequencing. Genome Res. 2011; 21(1):56-67.
13. Ren S, Peng Z, Mao JH, Yu Y, Yin C, Gao X, Cui Z, Zhang J, Yi K, Xu W, Chen C, Wang F, Guo X, Lu J, Yang J, Wei M, et al. RNA-seq analysis of prostate cancer in the Chinese population identifies recurrent gene fusions, cancer-associated long noncoding RNAs and aberrant alternative splicings. Cell Res. 2012; 22(5):806-821.
14. Kannan K, Wang L, Wang J, Ittmann MM, Li W and Yen L. Recurrent chimeric RNAs enriched in human prostate cancer identified by deep sequencing. Proc Natl Acad Sci U S A. 2011; 108(22):9172-9177.
15. Petrovics G, Zhang W, Makarem M, Street JP, Connelly R, Sun L, Sesterhenn IA, Srikantan V, Moul JW and Srivastava S. Elevated expression of PCGEM1, a prostate-specific gene with cell growth-promoting function, is associated with high-risk prostate cancer patients. Oncogene. 2004; 23(2):605-611.
16. Shilov IV, Seymour SL, Patel AA, Loboda A, Tang WH, Keating SP, Hunter CL, Nuwaysir LM and Schaeffer DA. The Paragon Algorithm, a next generation search engine that uses sequence temperature values and feature probabilities to identify peptides from tandem mass spectra. Mol Cell Proteomics. 2007; 6(9):1638-1655.
17. Kim JS, Monroe ME, Camp DG, 2nd, Smith RD and Qian WJ. In-source fragmentation and the sources of partially tryptic peptides in shotgun proteomics. J Proteome Res. 2013; 12(2):910-916.
18. Picotti P, Aebersold R and Domon B. The implications of proteolytic background for shotgun proteomics. Mol Cell Proteomics. 2007; 6(9):1589-1598.
19. Olsen JV, Ong SE and Mann M. Trypsin cleaves exclusively C-terminal to arginine and lysine residues. Mol Cell Proteomics. 2004; 3(6):608-614.
20. Prensner JR and Chinnaiyan AM. The emergence of lncRNAs in cancer biology. Cancer Discov. 2011; 1(5):391-407.
21. Kretz M, Siprashvili Z, Chu C, Webster DE, Zehnder A, Qu K, Lee CS, Flockhart RJ, Groff AF, Chow J, Johnston D, Kim GE, Spitale RC, Flynn RA, Zheng GX, Aiyer S, et al. Control of somatic tissue differentiation by the long non-coding RNA TINCR. Nature. 2013; 493(7431):231-235.

nature genetics

The long noncoding RNA *SChLAP1* promotes aggressive prostate cancer and antagonizes the SWI/SNF complex

John R Prensner[1,10], Matthew K Iyer[1,2,10], Anirban Sahu[1,10], Irfan A Asangani[1], Qi Cao[1], Lalit Patel[1,3], Ismael A Vergara[4], Elai Davicioni[4], Nicholas Erho[4], Mercedeh Ghadessi[4], Robert B Jenkins[5], Timothy J Triche[4], Rohit Malik[1], Rachel Bedenis[3], Natalie McGregor[3], Teng Ma[6], Wei Chen[6], Sumin Han[6], Xiaojun Jing[1], Xuhong Cao[1], Xiaoju Wang[1], Benjamin Chandler[1], Wei Yan[1], Javed Siddiqui[1], Lakshmi P Kunju[1,7,8], Saravana M Dhanasekaran[1,7], Kenneth J Pienta[1,3], Felix Y Feng[1,6,8] & Arul M Chinnaiyan[1,2,7–9]

Prostate cancers remain indolent in the majority of individuals but behave aggressively in a minority[1,2]. The molecular basis for this clinical heterogeneity remains incompletely understood[3–5]. Here we characterize a long noncoding RNA termed *SChLAP1* (second chromosome locus associated with prostate-1; also called *LINC00913*) that is overexpressed in a subset of prostate cancers. *SChLAP1* levels independently predict poor outcomes, including metastasis and prostate cancer–specific mortality. *In vitro* and *in vivo* gain-of-function and loss-of-function experiments indicate that *SChLAP1* is critical for cancer cell invasiveness and metastasis. Mechanistically, *SChLAP1* antagonizes the genome-wide localization and regulatory functions of the SWI/SNF chromatin-modifying complex. These results suggest that *SChLAP1* contributes to the development of lethal cancer at least in part by antagonizing the tumor-suppressive functions of the SWI/SNF complex.

With over 200,000 new cases per year, prostate cancer will be diagnosed in 1 in 6 men in the United States during their lifetime, yet only 20% of individuals with prostate cancer have a high-risk cancer that represents potentially lethal disease[1,2,4]. Whereas mutational events in key genes characterize a subset of lethal prostate cancers[3,5,6], the molecular basis for aggressive disease remains poorly understood.

Long noncoding RNAs (lncRNAs) are RNA species >200 bp in length that are frequently polyadenylated and associated with transcription by RNA polymerase II (ref. 7). lncRNA-mediated biology has been implicated in a wide variety of cellular processes, and, in cancer, lncRNAs are emerging as a prominent layer of transcriptional regulation, often by collaborating with epigenetic complexes[7–10].

Here we hypothesized that prostate cancer aggressiveness was governed by uncharacterized lncRNAs and sought to identify lncRNAs associated with aggressive disease. We previously used RNA sequencing (RNA-seq) to describe 121 new lncRNA loci (out of >1,800) that were aberrantly expressed in prostate cancer tissues[11]. Because only a fraction of prostate cancers present with aggressive clinical features[2], we performed cancer outlier profile analysis[11] (COPA) to nominate intergenic lncRNAs selectively upregulated in a subset of cancers (**Supplementary Table 1**). We observed that only two, *PCAT-109* and *PCAT-114*, which are both located in a 'gene desert' on chromosome 2q31.3 (**Supplementary Fig. 1**), had striking outlier profiles distinguishing them from the rest of the candidates[11] (**Fig. 1a**).

Of these two lncRNAs, *PCAT-114* was expressed at higher levels in prostate cell lines, and, in the *PCAT-114* region, we defined a 1.4-kb polyadenylated gene composed of up to seven exons and spanning nearly 200 kb on chromosome 2q31.3 (**Fig. 1b** and **Supplementary Fig. 2a**). We named this gene second chromosome locus associated with prostate-1 (*SChLAP1*) after its genomic location. Published prostate cancer chromatin immunoprecipitation and sequencing (ChIP-seq) data[12] confirmed that the transcriptional start site (TSS) of *SChLAP1* was marked by trimethylation of histone H3 at lysine 4 (H3K4me3) and that its gene body harbored trimethylation of histone H3 at lysine 36 (H3K36me3) (**Fig. 1b**), an epigenetic signature consistent with lncRNAs[13]. We observed numerous *SChLAP1* splicing isoforms, of which three (termed isoforms 1, 2 and 3) constituted the vast majority (>90%) of transcripts in the cell (**Supplementary Fig. 2b**,**c**).

Using quantitative PCR (qPCR), we confirmed that *SChLAP1* was highly expressed in ~25% of prostate cancers (**Fig. 1c**). *SChLAP1* was found to be expressed more frequently in metastatic compared to localized prostate cancers, and its expression was associated with *ETS* gene fusions in this cohort but not with other molecular events (**Supplementary Fig. 2d**,**e**). A computational analysis of the *SChLAP1* sequence suggested no coding potential, which was confirmed

[1]Michigan Center for Translational Pathology, University of Michigan, Ann Arbor, Michigan, USA. [2]Department of Computational Medicine and Bioinformatics, University of Michigan, Ann Arbor, Michigan, USA. [3]Department of Internal Medicine, University of Michigan, Ann Arbor, Michigan, USA. [4]GenomeDx Biosciences, Inc., Vancouver, British Columbia, Canada. [5]Department of Laboratory Medicine and Pathology, Mayo Clinic, Rochester, Minnesota, USA. [6]Department of Radiation Oncology, University of Michigan, Ann Arbor, Michigan, USA. [7]Department of Pathology, University of Michigan, Ann Arbor, Michigan, USA. [8]Comprehensive Cancer Center, University of Michigan, Ann Arbor, Michigan, USA. [9]Howard Hughes Medical Institute, University of Michigan, Ann Arbor, Michigan, USA. [10]These authors contributed equally to this work. Correspondence should be addressed to A.M.C. (arul@med.umich.edu).

Received 18 February; accepted 30 August; published online 29 September 2013; doi:10.1038/ng.2771

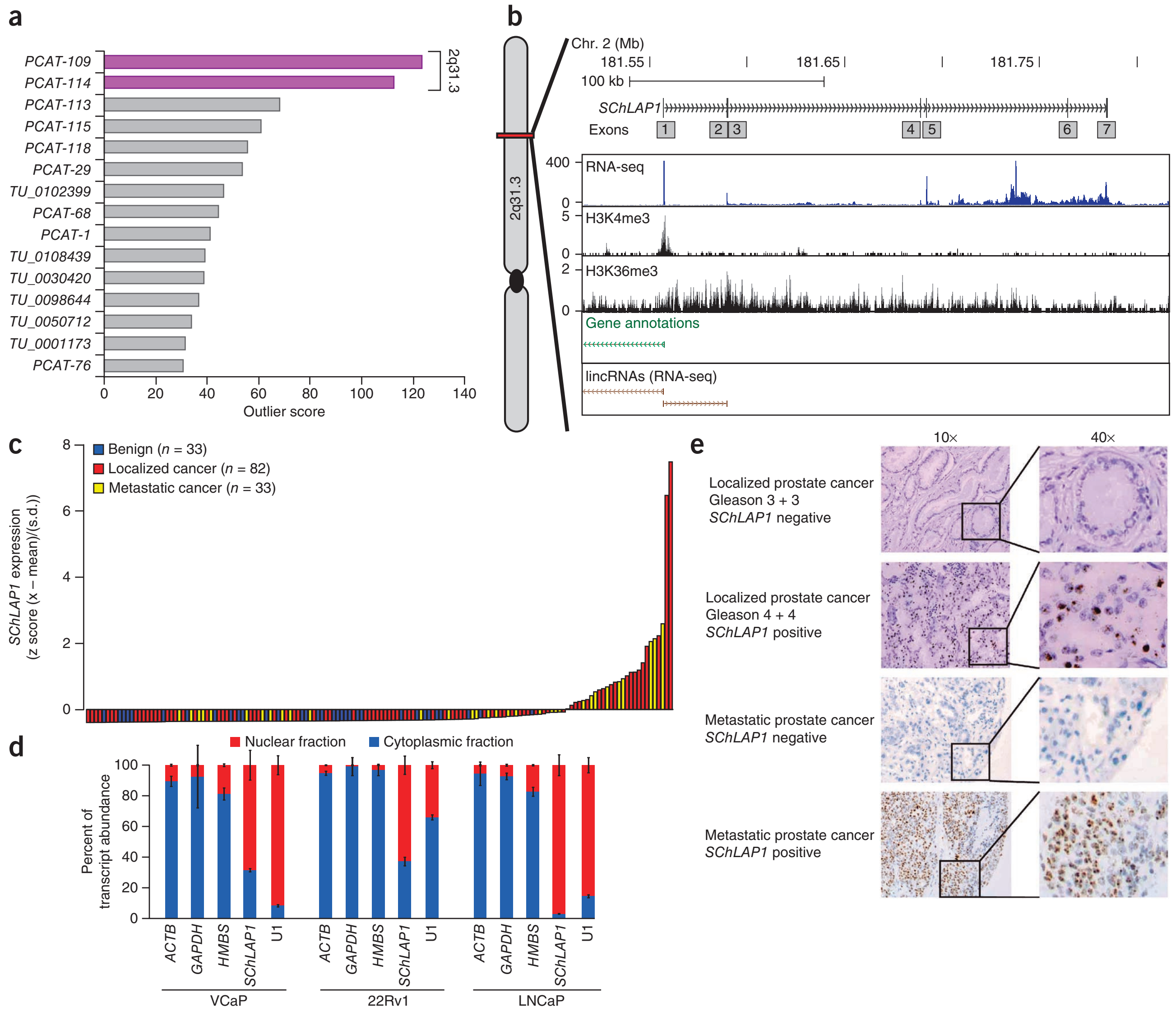


Figure 1 Identification of *SChLAP1* as a prostate cancer–associated lncRNA. (**a**) COPA for intergenic lncRNAs (lncRNAs defined in ref. 11). (**b**) Representation of the *SChLAP1* gene and its annotations in current databases. An aggregated representation of current gene annotations for Ensembl, the Encyclopedia of DNA Elements (ENCODE), the UCSC Genome Browser, RefSeq and Vega shows no annotation for *SChLAP1*. ChIP-seq data for H3K4me3 and H3K36me3 show enrichment at the *SChLAP1* gene. Also, RNA-seq data showing an outlier sample for *SChLAP1* demonstrate its expression. (**c**) qPCR for *SChLAP1* on a panel of benign prostate (*n* = 33), localized prostate cancer (*n* = 82) and metastatic prostate cancer (*n* = 33) samples. qPCR data are normalized to the average of (*GAPDH* + *HMBS*) and are represented as standardized expression values. (**d**) Fractionation of prostate cell lysates demonstrates nuclear expression of *SChLAP1*. U1 RNA serves as a positive control for nuclear gene expression. Error bars, s.e.m. (**e**) *In situ* hybridization of *SChLAP1* in human prostate cancer. Histological scores for localized cancer samples are indicated, with the first number representing the major Gleason score and the second number representing the minor Gleason score. *SChLAP1* staining is shown for both localized and metastatic tissues.

experimentally by *in vitro* translation assays of the three *SChLAP1* isoforms (**Supplementary Fig. 3**). Additionally, we found that *SChLAP1* transcripts were located in the nucleus (**Fig. 1d**). We confirmed the nuclear localization of *SChLAP1* transcripts in human samples (**Fig. 1e**) using an *in situ* hybridization assay in formalin-fixed, paraffin-embedded prostate cancer samples (**Supplementary Fig. 4a**,**b** and **Supplementary Note**).

An analysis of *SChLAP1* expression in localized tumors demonstrated a strong correlation with higher Gleason scores, a histopathological measure of aggressiveness (**Supplementary Fig. 4c**,**d** and **Supplementary Table 2**). Next, we performed a network analysis of prostate cancer microarray data in the Oncomine[14] database using signatures of *SChLAP1*-correlated or *SChLAP1*-anticorrelated genes, as *SChLAP1* itself is not measured by expression microarrays (Online Methods and **Supplementary Table 3a**). We found a striking association with enriched concepts related to prostate cancer progression (**Fig. 2a** and **Supplementary Table 3b**). For comparison, we next incorporated disease signatures using prostate RNA-seq data and additional known prostate cancer genes, including *EZH2* (a metastasis gene[15]), *PCA3* (a lncRNA biomarker[4]) and *AMACR* (a tissue biomarker[4]), as well as *ACTB* (encoding β-actin) as a control (**Supplementary Fig. 5**, **Supplementary Table 3c–i** and **Supplementary Note**). A heatmap visualization of significant comparisons confirmed a strong association of *SChLAP1*-correlated genes but not of

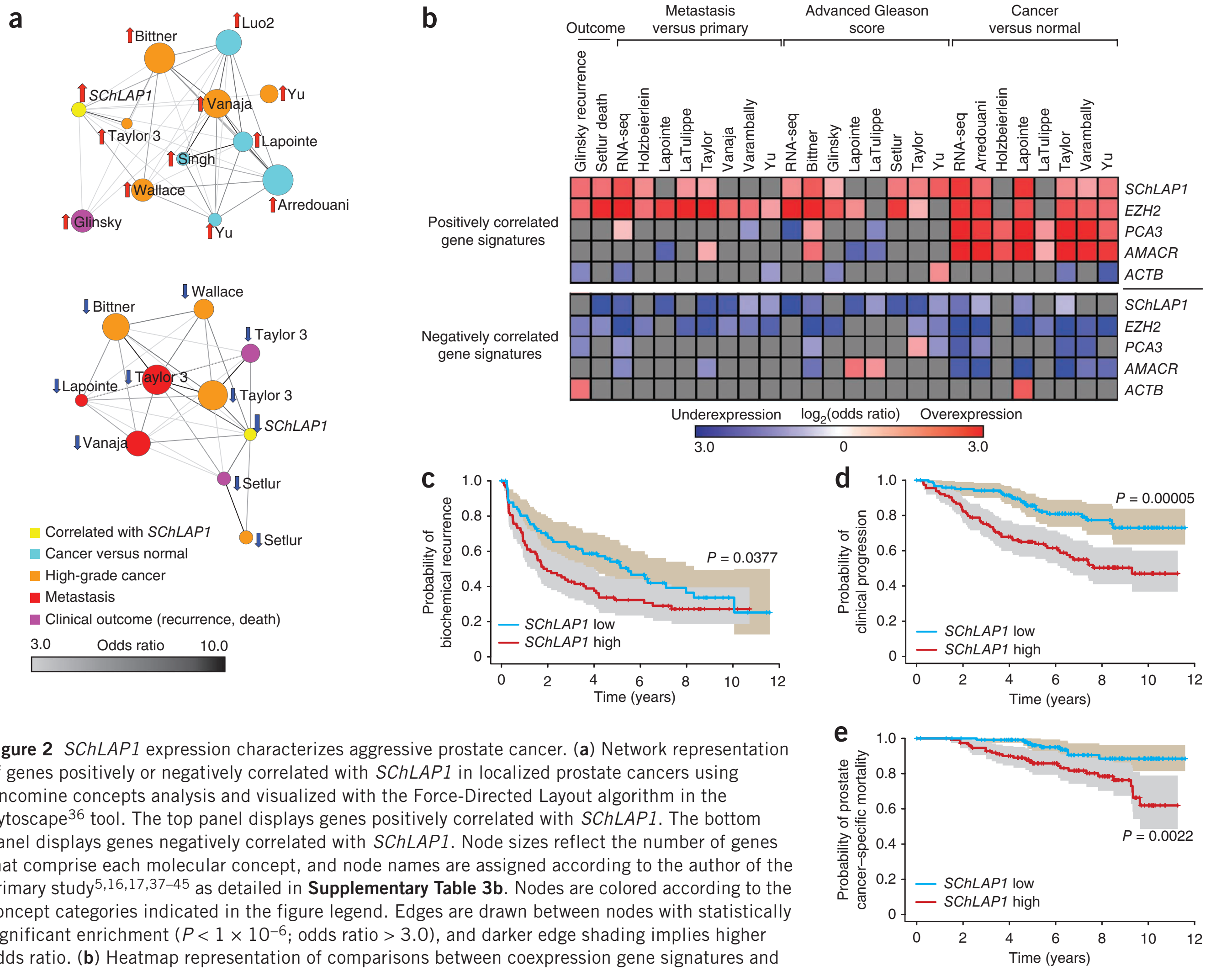


Figure 2 *SChLAP1* expression characterizes aggressive prostate cancer. (**a**) Network representation of genes positively or negatively correlated with *SChLAP1* in localized prostate cancers using Oncomine concepts analysis and visualized with the Force-Directed Layout algorithm in the Cytoscape[36] tool. The top panel displays genes positively correlated with *SChLAP1*. The bottom panel displays genes negatively correlated with *SChLAP1*. Node sizes reflect the number of genes that comprise each molecular concept, and node names are assigned according to the author of the primary study[5,16,17,37–45] as detailed in **Supplementary Table 3b**. Nodes are colored according to the concept categories indicated in the figure legend. Edges are drawn between nodes with statistically significant enrichment ($P < 1 \times 10^{-6}$; odds ratio > 3.0), and darker edge shading implies higher odds ratio. (**b**) Heatmap representation of comparisons between coexpression gene signatures and molecular concepts. Comparisons to positively (top) and negatively (bottom) correlated gene signatures are shown separately. Comparisons that do not reach statistical significance ($q > 0.01$ or odds ratio < 2) are shown in gray. Associations with overexpression concepts are colored red, and underexpression concepts are colored blue. (**c**–**e**) Kaplan-Meier analyses of prostate cancer outcomes in the Mayo Clinic cohort. *SChLAP1* expression was measured using Affymetrix exon arrays, and subjects were stratified according to their *SChLAP1* expression level. Subject outcomes were analyzed for biochemical recurrence (**c**), clinical progression to systemic disease (**d**) and prostate cancer–specific mortality (**e**). The shaded regions represent 95% confidence intervals. *P* values for Kaplan-Meier curves were determined using a log-rank test.

PCA3- and *AMACR*-correlated genes with high-grade and metastatic cancers (**Fig. 2b**). Kaplan-Meier analysis similarly showed significant associations between the *SChLAP1* signature and biochemical recurrence[16] and overall survival[17] (**Supplementary Fig. 6a,b**).

To directly evaluate the relationship between *SChLAP1* levels and clinical outcome, we next used *SChLAP1* expression to stratify 235 samples from individuals with localized prostate cancer who underwent radical prostatectomy at the Mayo Clinic[18] (Online Methods and **Supplementary Fig. 6c**). We evaluated samples for three clinical endpoints: biochemical recurrence, clinical progression to systemic disease and prostate cancer–specific mortality (**Supplementary Table 4**). At the time of this analysis, subjects had a median follow-up time of 8.1 years.

SChLAP1 was a powerful single-gene predictor of aggressive prostate cancer (**Fig. 2c**–**e**). *SChLAP1* expression was highly significant when distinguishing disease with clinical progression and prostate cancer–specific mortality (P = 0.00005 and 0.002, respectively; **Fig. 2d,e**). For the biochemical recurrence endpoint, high *SChLAP1* expression was associated with a shorter median time to progression (1.9 versus 5.5 years for individuals with high and low expression of *SChLAP1*, respectively; **Fig. 2c**). We further confirmed this association with rapid biochemical recurrence using an independent cohort (**Supplementary Fig. 6d**). Multivariate and univariate regression analyses of the Mayo Clinic data demonstrated that *SChLAP1* expression is an independent predictor of prostate cancer aggressiveness, with highly significant hazard ratios for predicting biochemical recurrence, clinical progression and prostate cancer–specific mortality (hazard ratios of 3.045, 3.563 and 4.339, respectively; $P < 0.01$), which are comparable to those for other clinical factors such as advanced clinical stage and Gleason histopathological score (**Supplementary Fig. 7** and **Supplementary Note**).

To explore the functional role of *SChLAP1*, we performed small interfering RNA (siRNA)-mediated knockdowns to compare the impact of *SChLAP1* depletion to that of *EZH2*, which is essential for cancer cell aggressiveness[15]. Notably, knockdown of *SChLAP1* dramatically impaired cell invasion and proliferation *in vitro* to

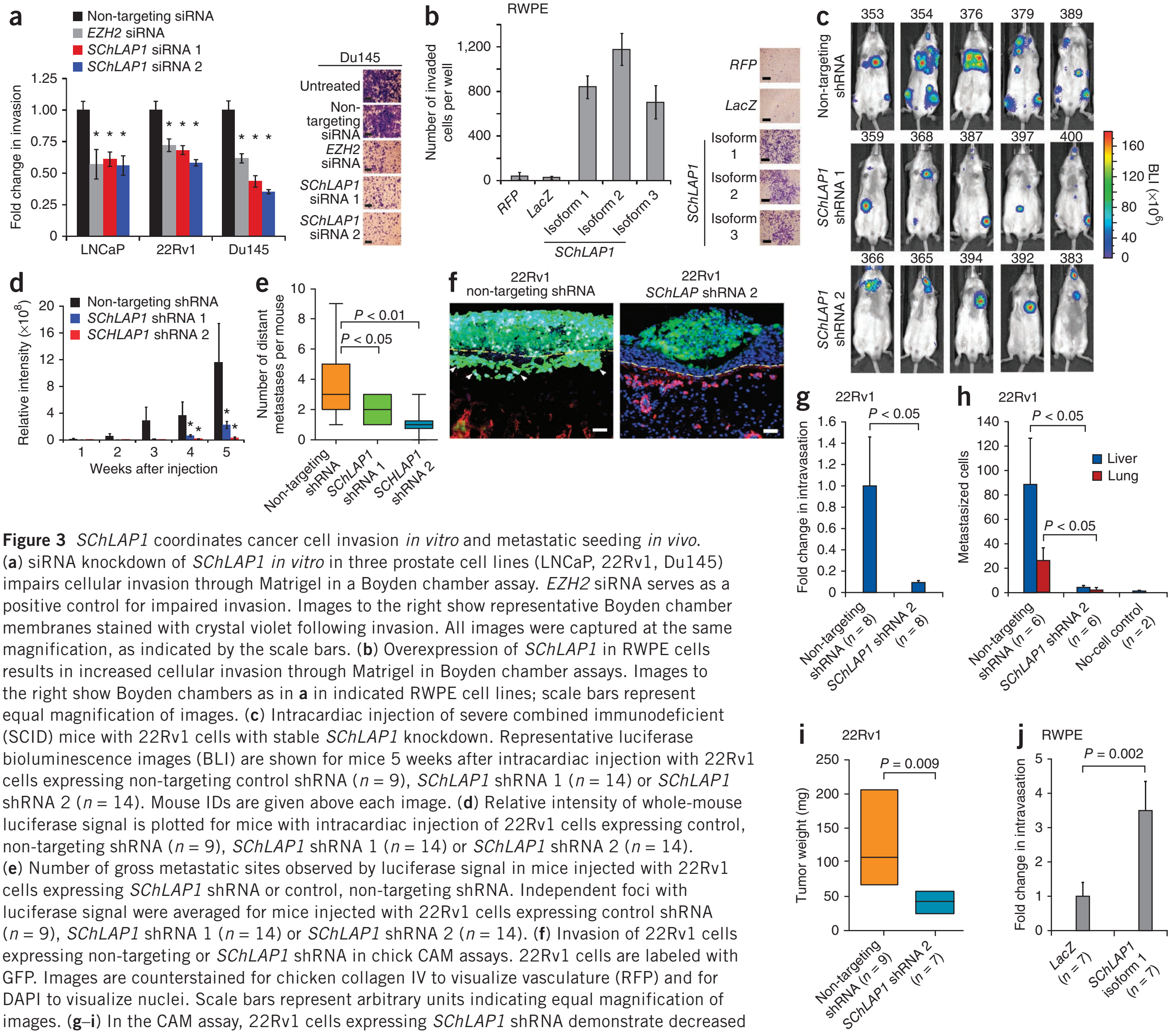


Figure 3 *SChLAP1* coordinates cancer cell invasion *in vitro* and metastatic seeding *in vivo*. (**a**) siRNA knockdown of *SChLAP1 in vitro* in three prostate cell lines (LNCaP, 22Rv1, Du145) impairs cellular invasion through Matrigel in a Boyden chamber assay. *EZH2* siRNA serves as a positive control for impaired invasion. Images to the right show representative Boyden chamber membranes stained with crystal violet following invasion. All images were captured at the same magnification, as indicated by the scale bars. (**b**) Overexpression of *SChLAP1* in RWPE cells results in increased cellular invasion through Matrigel in Boyden chamber assays. Images to the right show Boyden chambers as in **a** in indicated RWPE cell lines; scale bars represent equal magnification of images. (**c**) Intracardiac injection of severe combined immunodeficient (SCID) mice with 22Rv1 cells with stable *SChLAP1* knockdown. Representative luciferase bioluminescence images (BLI) are shown for mice 5 weeks after intracardiac injection with 22Rv1 cells expressing non-targeting control shRNA ($n = 9$), *SChLAP1* shRNA 1 ($n = 14$) or *SChLAP1* shRNA 2 ($n = 14$). Mouse IDs are given above each image. (**d**) Relative intensity of whole-mouse luciferase signal is plotted for mice with intracardiac injection of 22Rv1 cells expressing control, non-targeting shRNA ($n = 9$), *SChLAP1* shRNA 1 ($n = 14$) or *SChLAP1* shRNA 2 ($n = 14$). (**e**) Number of gross metastatic sites observed by luciferase signal in mice injected with 22Rv1 cells expressing *SChLAP1* shRNA or control, non-targeting shRNA. Independent foci with luciferase signal were averaged for mice injected with 22Rv1 cells expressing control shRNA ($n = 9$), *SChLAP1* shRNA 1 ($n = 14$) or *SChLAP1* shRNA 2 ($n = 14$). (**f**) Invasion of 22Rv1 cells expressing non-targeting or *SChLAP1* shRNA in chick CAM assays. 22Rv1 cells are labeled with GFP. Images are counterstained for chicken collagen IV to visualize vasculature (RFP) and for DAPI to visualize nuclei. Scale bars represent arbitrary units indicating equal magnification of images. (**g**–**i**) In the CAM assay, 22Rv1 cells expressing *SChLAP1* shRNA demonstrate decreased intravasation (**g**), metastatic spread to the liver and lungs (**h**) and reduced tumor weight (**i**) relative to 22Rv1 cells expressing control, non-targeting shRNA. (**j**) Quantification of intravasation of RWPE cells expressing *LacZ* or *SChLAP1* in the CAM assay. All data in bar plots are represented as mean ± s.e.m. Statistical significance was determined by two-tailed Student's *t* test: $*P < 0.05$. Box plots in **e**,**i** display box-and-whisker plots with the midpoint line indicating the median, box boundaries showing 25th and 75th quartile ranges and whiskers displaying the minimum and maximum values.

an extent comparable to that observed with knockdown of *EZH2* (**Fig. 3a** and **Supplementary Fig. 8a**,**b**). Overexpression of an siRNA-resistant *SChLAP1* isoform rescued the *in vitro* invasive phenotype of 22Rv1 cells treated with siRNA-2 (**Supplementary Fig. 8c**,**d**). Overexpression of the three *SChLAP1* isoforms in benign, immortalized RWPE prostate cells dramatically increased the ability of these cells to invade *in vitro* but did not affect cell proliferation (**Fig. 3b** and **Supplementary Fig. 8e**,**f**).

To test *SChLAP1 in vivo*, we performed intracardiac injection of CB-17 SCID mice with 22Rv1 cells stably knocking down *SChLAP1* (**Supplementary Fig. 9a**) and observed that *SChLAP1* depletion impaired metastatic seeding and growth, as measured by luciferase signaling at both proximal (lungs) and distal sites (**Fig. 3c**,**d**). Indeed, compared to mice injected with 22Rv1 cells expressing a non-targeting control, mice injected with 22Rv1 cells stably expressing short hairpin RNA (shRNA) against *SChLAP1* had both fewer gross metastatic sites overall as well as smaller metastatic tumors when they did form (**Fig. 3d**,**e**). Histopathological analysis of the metastatic 22Rv1 tumors, regardless of *SChLAP1* knockdown, showed uniformly high-grade epithelial cancer (**Supplementary Fig. 9b**). Interestingly, subcutaneous xenografts with stable knockdown of *SChLAP1* showed slower tumor progression; however, this was due to delayed tumor engraftment rather than to decreased tumor growth kinetics, with no change in Ki67 staining observed between cells expressing *SChLAP1* shRNA and control cells expressing non-targeting shRNA (**Supplementary Fig. 9c**–**i**).

Next, using the chick chorioallantoic membrane (CAM) assay[19], we found that 22Rv1 cells expressing *SChLAP1* shRNA 2, which have depleted expression of both isoforms 1 and 2, had greatly reduced ability to invade, intravasate and metastasize to distant organs (**Fig. 3f**–**h**). Additionally, cells with knockdown of *SChLAP1* also

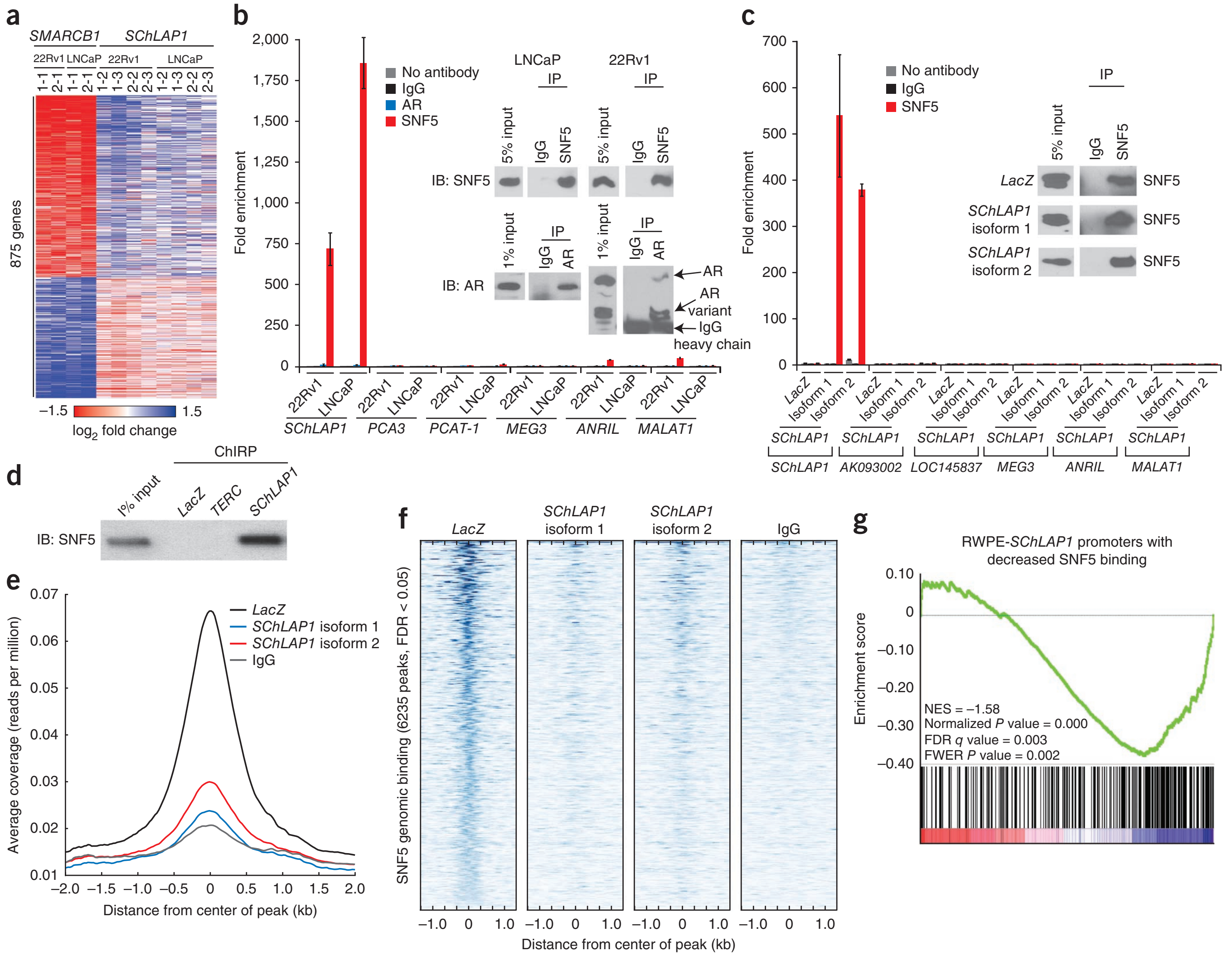


Figure 4 *SChLAP1* antagonizes SNF5 function and attenuates SNF5 genome-wide localization. (**a**) Heatmap results for *SChLAP1* or *SMARCB1* knockdown in LNCaP and 22Rv1 cells. The numbers above the heatmap indicate the specific siRNA and microarray replicates. (**b**) RIP of SNF5 or androgen receptor (AR) in 22Rv1 and LNCaP cells. Inset, protein blots showing pulldown efficiency. Error bars, s.e.m. (**c**) RIP analysis of SNF5 in RWPE cells overexpressing *LacZ*, *SChLAP1* isoform 1 or *SChLAP1* isoform 2. Inset, protein blots showing pulldown efficiency. Error bars, s.e.m. (**d**) Pulldown of *SChLAP1* RNA using chromatin isolation by RNA purification (ChIRP) recovers SNF5 protein in RWPE cells expressing *SChLAP1* isoform 1. *LacZ* and *TERC* serve as negative and positive controls, respectively. (**e**) Global representation of SNF5 genomic binding over a 4-kb window centered on each SNF5 ChIP-seq peak in RWPE cells expressing *LacZ*, *SChLAP1* isoform 1 or *SChLAP1* isoform 2. (**f**) Heatmap of SNF5 genomic binding at target sites in RWPE cells expressing *LacZ*, *SChLAP1* isoform 1 or *SChLAP1* isoform 2. A 2-kb interval centered on the called SNF5 peak is shown. (**g**) GSEA results showing significant enrichment of ChIP-seq promoter peaks with >2-fold loss of SNF5 binding for underexpressed genes in RWPE cells expressing *SChLAP1*. NES, normalized enrichment score; FWER, familywise error rate.

resulted in decreased tumor growth (**Fig. 3i**). Notably, RWPE cells with overexpression of *SChLAP1* isoform 1 partially supported these results, showing a markedly increased ability to intravasate (**Fig. 3j**). RWPE cells overexpressing *SChLAP1* did not generate distant metastases or cause altered tumor growth in this model (data not shown). Together, the mouse metastasis and CAM data strongly implicate *SChLAP1* in tumor invasion and metastasis through activity in cancer cell intravasation, extravasation and subsequent tumor cell seeding.

To elucidate the mechanisms of *SChLAP1* function, we profiled 22Rv1 and LNCaP cells with *SChLAP1* knockdown, identifying 165 upregulated and 264 downregulated genes (q value < 0.001) (**Supplementary Fig. 10a** and **Supplementary Table 5a**). After ranking genes according to differential expression[20], we employed Gene Set Enrichment Analysis (GSEA)[21] to search for enrichment across the Molecular Signatures Database (MSigDB)[22]. Among the highest ranked concepts, we noticed genes positively or negatively correlated with the SWI/SNF complex[23], and this association was independently confirmed using gene signatures generated from our RNA-seq data (**Supplementary Fig. 10b**–**e** and **Supplementary Table 5b**,**c**). Notably, *SChLAP1*-regulated genes were inversely correlated with these data sets, suggesting that *SChLAP1* and the SWI/SNF complex function in opposing manners.

The SWI/SNF complex regulates gene transcription as a multiprotein system that physically moves nucleosomes at gene promoters[24]. Loss of SWI/SNF complex functionality promotes cancer progression, and multiple SWI/SNF components are somatically inactivated in cancer[24,25]. SWI/SNF complex mutations do occur in prostate cancer, albeit not commonly[3], and downregulation of SWI/SNF complex members characterizes subsets of prostate cancer[23,26]. Thus, antagonism of SWI/SNF complex activity by *SChLAP1* is consistent with the oncogenic behavior of *SChLAP1* and the tumor suppressive behavior of the SWI/SNF complex.

To directly test whether *SChLAP1* antagonizes SWI/SNF-mediated regulation, we performed siRNA-mediated knockdown of *SMARCB1* (which encodes the SNF5 protein) (**Supplementary Fig. 10f**), an essential subunit that facilitates SWI/SNF complex binding to histone proteins[24,25,27], and confirmed predicted expression changes for several *SChLAP1*- or SNF5-regulated genes (**Supplementary Fig. 10g,h**). A comparison of genes whose expression was altered by knockdown of *SMARCB1* to those regulated by *SChLAP1* demonstrated an antagonistic relationship in which *SChLAP1* knockdown affected the same genes as *SMARCB1* knockdown but with opposing directions of effect (**Fig. 4a** and **Supplementary Table 5d–h**). We used GSEA to quantify and verify the significance of these findings (false discovery rate (FDR) < 0.05) (**Supplementary Fig. 10i–k**). Furthermore, a shared *SMARCB1-SChLAP1* signature of coregulated genes was highly enriched for prostate cancer clinical signatures for disease aggressiveness (**Supplementary Fig. 11** and **Supplementary Table 5i**).

Mechanistically, although *SChLAP1* and *SMARCB1* mRNA levels were comparable (**Supplementary Fig. 12a**), *SChLAP1* knockdown or overexpression did not alter SNF5 protein abundance (**Supplementary Fig. 12b**), suggesting that *SChLAP1* regulates SWI/SNF activity post-translationally. To explore this possibility, we performed RNA immunoprecipitation assays (RIPs) for SNF5. We found that endogenous *SChLAP1* but not other cytoplasmic or nuclear lncRNAs[7,28] robustly coimmunoprecipitated with SNF5 under native conditions (**Fig. 4b**) and with use of UV cross-linking (**Supplementary Fig. 12c**), and coimmunoprecipitation was also observed with a second antibody to SNF5 (**Supplementary Fig. 12d**). In contrast, *SChLAP1* did not coimmunoprecipitate with androgen receptor (**Fig. 4b**). Furthermore, both *SChLAP1* isoform 1 and isoform 2 coimmunoprecipitated with SNF5 in RWPE overexpression models (**Fig. 4c** and **Supplementary Fig. 12e**). SNRNP70 binding to U1 RNA was used as a technical control in all cell lines (**Supplementary Fig. 12f,g**). Finally, pulldown of *SChLAP1* RNA in RWPE cells overexpressing *SChLAP1* isoform 1 robustly recovered SNF5 protein, confirming this interaction (**Fig. 4d** and **Supplementary Fig. 12h**).

To address whether *SChLAP1* modulates SWI/SNF genomic binding, we performed ChIP-seq for SNF5 in RWPE cells expressing *LacZ* or *SChLAP1* and called significantly enriched peaks with respect to an IgG control (Online Methods and **Supplementary Table 6a**). Protein blot validation confirmed SNF5 pulldown by ChIP (**Supplementary Fig. 13a**). After aggregating called peaks from all samples, we found 6,235 genome-wide binding sites for SNF5 (FDR < 0.05; **Supplementary Table 6b**), which were highly enriched for sites near gene promoters (**Supplementary Fig. 13b**), supporting results from previous studies of SWI/SNF binding[29–31].

A comparison of SNF5 binding across these 6,235 genomic sites demonstrated a dramatic decrease in SNF5 genomic binding as a result of *SChLAP1* overexpression (**Fig. 4e,f** and **Supplementary Fig. 13c**). Of the 1,299 SNF5 peaks occurring within 1 kb of a gene TSS, 390 showed relative SNF5 binding that was decreased by ≥2-fold with *SChLAP1* overexpression (**Supplementary Fig. 13d** and **Supplementary Table 6c**). To verify these findings independently, we performed ChIP for SNF5 in 22Rv1 cells expressing shRNA to *SChLAP1*, with the hypothesis that knockdown of *SChLAP1* should increase SNF5 genomic binding compared to controls. We found that 9 of 12 target genes showed a substantial increase in SNF5 binding with knockdown of *SChLAP1* (**Supplementary Fig. 14a**), confirming our predictions.

Finally, we used expression profiling of RWPE cells expressing *LacZ* or *SChLAP1* to characterize the relationship between SNF5 binding and *SChLAP1*-mediated changes in gene expression. After identifying a gene signature with highly significant changes in expression (**Supplementary Table 6d**), we intersected this signature with the ChIP-seq data. We observed that a substantial subset of genes with ≥2-fold relative decrease in SNF5 genomic binding were dysregulated when *SChLAP1* was overexpressed (**Supplementary Fig. 14b**). Decreased SNF5 binding was primarily associated with the downregulation of target gene expression (**Supplementary Table 6e**), although the SWI/SNF complex is known to regulate expression in either direction[24,25]. Integrative GSEA of the microarray and SNF5 ChIP-seq data demonstrated significant enrichment for genes that were repressed when *SChLAP1* was overexpressed (q value = 0.003; **Fig. 4g**). Overall, these data argue that *SChLAP1* overexpression antagonizes SWI/SNF complex function by attenuating the genomic binding of this complex, thereby impairing its ability to properly regulate gene expression.

Here we have discovered *SChLAP1*, a highly prognostic lncRNA that is abundantly expressed in ~25% of prostate cancers and that aids in the discrimination of aggressive tumors from indolent forms of the disease. Mechanistically, we find that *SChLAP1* coordinates cancer cell invasion *in vitro* and metastatic spread *in vivo*. Moreover, we characterize an antagonistic *SChLAP1*-SWI/SNF axis in which *SChLAP1* impairs SNF5-mediated regulation of gene expression and genomic binding (**Supplementary Fig. 14c**). Thus, whereas other lncRNAs such as *HOTAIR* and *HOTTIP* are known to assist epigenetic complexes such as PRC2 and MLL by facilitating their genomic binding and enhancing their functions[8,9,32], *SChLAP1* is the first lncRNA, to our knowledge, that impairs a major epigenetic complex with well-documented tumor suppressor function[23–25,33–35]. Our discovery of *SChLAP1* has broad implications for cancer biology and provides supporting evidence for the role of lncRNAs in the progression of aggressive cancers.

URLs. Stellaris probe designer, http://www.singlemoleculefish.com; HT-Seq, http://www-huber.embl.de/users/anders/HTSeq/; BioVenn, http://www.cmbi.ru.nl/cdd/biovenn/; Galaxy, http://usegalaxy.org/.

METHODS

Methods and any associated references are available in the online version of the paper.

Accession codes. Sequences for *SChLAP1* isoforms 1–7 have been deposited in GenBank under accessions JX117418, JX117419, JX117420, JX117421, JX117422, JX117423 and JX117424. Microarray data have been deposited in the Gene Expression Omnibus (GEO) under accession GSE40386.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS

We thank O.A. Balbin, S.A. Tomlins, C. Brenner, S. Deroo and S. Roychowdhury for helpful discussions. This work was supported in part by US National Institutes of Health (NIH) Prostate Specialized Program of Research Excellence grant P50CA69568, Early Detection Research Network grant UO1 CA111275, US NIH grant R01CA132874-01A1 and US Department of Defense grant PC100171 (A.M.C.). A.M.C. is supported by a Doris Duke Charitable Foundation Clinical Scientist Award, by the Prostate Cancer Foundation and by the Howard Hughes Medical Institute. A.M.C. is an American Cancer Society Research Professor. A.M.C. is a Taubman Scholar of the University of Michigan. F.Y.F. was supported by the Prostate Cancer Foundation and by US Department of Defense grant PC094231. Q.C. was supported by US Department of Defense Postdoctoral

Fellowship PC094725. J.R.P. was supported by US Department of Defense Predoctoral Fellowship PC094290. M.K.I. was supported by US Department of Defense Predoctoral Fellowship BC100238. A.S. was supported by NIH Predoctoral Fellowship 1F30CA180376-01. J.R.P., M.K.I. and A.S. are Fellows of the University of Michigan Medical Scientist Training Program.

AUTHOR CONTRIBUTIONS

J.R.P., M.K.I., A.S. and A.M.C. designed the project and directed experimental studies. J.R.P., Q.C., W.C., S.M.D., B.C., S.H., R.M., L.P., T.M. and A.S. performed *in vitro* studies. X.W. performed *in vitro* translation assays. I.A.A. and A.S. performed CAM assays. R.B., N.M. and K.J.P. performed *in vivo* studies. L.P.K. and W.Y. performed histopathological analyses. M.K.I. performed bioinformatics analysis. X.J. and X.C. performed gene expression microarray experiments. J.S. and F.Y.F. facilitated biological sample procurement. F.Y.F. performed clinical analyses. For the Mayo Clinic cohort, R.B.J. provided clinical samples and outcomes data. T.J.T. and E.D. generated and analyzed expression profiles for the Mayo Clinic cohort. E.D., N.E., M.G. and I.A.V. performed statistical analyses of *SChLAP1* expression in the Mayo Clinic cohort. J.R.P., M.K.I., A.S. and A.M.C. interpreted data and wrote the manuscript.

COMPETING FINANCIAL INTERESTS

The authors declare competing financial interests: details are available in the online version of the paper.

Reprints and permissions information is available online at http://www.nature.com/reprints/index.html.

1. Etzioni, R., Cha, R., Feuer, E.J. & Davidov, O. Asymptomatic incidence and duration of prostate cancer. *Am. J. Epidemiol.* **148**, 775–785 (1998).
2. Cooperberg, M.R., Moul, J.W. & Carroll, P.R. The changing face of prostate cancer. *J. Clin. Oncol.* **23**, 8146–8151 (2005).
3. Grasso, C.S. *et al.* The mutational landscape of lethal castration-resistant prostate cancer. *Nature* **487**, 239–243 (2012).
4. Prensner, J.R., Rubin, M.A., Wei, J.T. & Chinnaiyan, A.M. Beyond PSA: the next generation of prostate cancer biomarkers. *Sci. Transl. Med.* **4**, 127rv3 (2012).
5. Taylor, B.S. *et al.* Integrative genomic profiling of human prostate cancer. *Cancer Cell* **18**, 11–22 (2010).
6. Berger, M.F. *et al.* The genomic complexity of primary human prostate cancer. *Nature* **470**, 214–220 (2011).
7. Prensner, J.R. & Chinnaiyan, A.M. The emergence of lncRNAs in cancer biology. *Cancer Discov.* **1**, 391–407 (2011).
8. Rinn, J.L. *et al.* Functional demarcation of active and silent chromatin domains in human *HOX* loci by noncoding RNAs. *Cell* **129**, 1311–1323 (2007).
9. Tsai, M.C. *et al.* Long noncoding RNA as modular scaffold of histone modification complexes. *Science* **329**, 689–693 (2010).
10. Kotake, Y. *et al.* Long non-coding RNA *ANRIL* is required for the PRC2 recruitment to and silencing of *p15*^{INK4B} tumor suppressor gene. *Oncogene* **30**, 1956–1962 (2011).
11. Prensner, J.R. *et al.* Transcriptome sequencing across a prostate cancer cohort identifies *PCAT-1*, an unannotated lincRNA implicated in disease progression. *Nat. Biotechnol.* **29**, 742–749 (2011).
12. Yu, J. *et al.* An integrated network of androgen receptor, polycomb, and *TMPRSS2-ERG* gene fusions in prostate cancer progression. *Cancer Cell* **17**, 443–454 (2010).
13. Guttman, M. *et al.* Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. *Nature* **458**, 223–227 (2009).
14. Rhodes, D.R. *et al.* Oncomine 3.0: genes, pathways, and networks in a collection of 18,000 cancer gene expression profiles. *Neoplasia* **9**, 166–180 (2007).
15. Varambally, S. *et al.* The polycomb group protein EZH2 is involved in progression of prostate cancer. *Nature* **419**, 624–629 (2002).
16. Glinsky, G.V., Glinskii, A.B., Stephenson, A.J., Hoffman, R.M. & Gerald, W.L. Gene expression profiling predicts clinical outcome of prostate cancer. *J. Clin. Invest.* **113**, 913–923 (2004).
17. Setlur, S.R. *et al.* Estrogen-dependent signaling in a molecularly distinct subclass of aggressive prostate cancer. *J. Natl. Cancer Inst.* **100**, 815–825 (2008).
18. Nakagawa, T. *et al.* A tissue biomarker panel predicting systemic progression after PSA recurrence post-definitive prostate cancer therapy. *PLoS ONE* **3**, e2318 (2008).
19. Asangani, I.A. *et al.* Characterization of the EZH2-MMSET histone methyltransferase regulatory axis in cancer. *Mol. Cell* **49**, 80–93 (2013).
20. Tusher, V.G., Tibshirani, R. & Chu, G. Significance analysis of microarrays applied to the ionizing radiation response. *Proc. Natl. Acad. Sci. USA* **98**, 5116–5121 (2001).
21. Subramanian, A. *et al.* Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. USA* **102**, 15545–15550 (2005).
22. Liberzon, A. *et al.* Molecular signatures database (MSigDB) 3.0. *Bioinformatics* **27**, 1739–1740 (2011).
23. Shen, H. *et al.* The SWI/SNF ATPase Brm is a gatekeeper of proliferative control in prostate cancer. *Cancer Res.* **68**, 10154–10162 (2008).
24. Roberts, C.W. & Orkin, S.H. The SWI/SNF complex—chromatin and cancer. *Nat. Rev. Cancer* **4**, 133–142 (2004).
25. Reisman, D., Glaros, S. & Thompson, E.A. The SWI/SNF complex and cancer. *Oncogene* **28**, 1653–1668 (2009).
26. Sun, A. *et al.* Aberrant expression of SWI/SNF catalytic subunits BRG1/BRM is associated with tumor development and increased invasiveness in prostate cancers. *Prostate* **67**, 203–213 (2007).
27. Dechassa, M.L. *et al.* Architecture of the SWI/SNF-nucleosome complex. *Mol. Cell. Biol.* **28**, 6010–6021 (2008).
28. Derrien, T. *et al.* The GENCODE v7 catalog of human long noncoding RNAs: analysis of their gene structure, evolution, and expression. *Genome Res.* **22**, 1775–1789 (2012).
29. De, S. *et al.* Dynamic BRG1 recruitment during T helper differentiation and activation reveals distal regulatory elements. *Mol. Cell. Biol.* **31**, 1512–1527 (2011).
30. Euskirchen, G.M. *et al.* Diverse roles and interactions of the SWI/SNF chromatin remodeling complex revealed using global approaches. *PLoS Genet.* **7**, e1002008 (2011).
31. Yen, K., Vinayachandran, V., Batta, K., Koerber, R.T. & Pugh, B.F. Genome-wide nucleosome specificity and directionality of chromatin remodelers. *Cell* **149**, 1461–1473 (2012).
32. Gupta, R.A. *et al.* Long non-coding RNA *HOTAIR* reprograms chromatin state to promote cancer metastasis. *Nature* **464**, 1071–1076 (2010).
33. Jones, S. *et al.* Frequent mutations of chromatin remodeling gene *ARID1A* in ovarian clear cell carcinoma. *Science* **330**, 228–231 (2010).
34. Varela, I. *et al.* Exome sequencing identifies frequent mutation of the SWI/SNF complex gene *PBRM1* in renal carcinoma. *Nature* **469**, 539–542 (2011).
35. Versteege, I. *et al.* Truncating mutations of hSNF5/INI1 in aggressive paediatric cancer. *Nature* **394**, 203–206 (1998).
36. Cline, M.S. *et al.* Integration of biological networks and gene expression data using Cytoscape. *Nat. Protoc.* **2**, 2366–2382 (2007).
37. Arredouani, M.S. *et al.* Identification of the transcription factor single-minded homologue 2 as a potential biomarker and immunotherapy target in prostate cancer. *Clin. Cancer Res.* **15**, 5794–5802 (2009).
38. Holzbeierlein, J. *et al.* Gene expression analysis of human prostate carcinoma during hormonal therapy identifies androgen-responsive genes and mechanisms of therapy resistance. *Am. J. Pathol.* **164**, 217–227 (2004).
39. Lapointe, J. *et al.* Gene expression profiling identifies clinically relevant subtypes of prostate cancer. *Proc. Natl. Acad. Sci. USA* **101**, 811–816 (2004).
40. LaTulippe, E. *et al.* Comprehensive gene expression analysis of prostate cancer reveals distinct transcriptional programs associated with metastatic disease. *Cancer Res.* **62**, 4499–4506 (2002).
41. Luo, J.H. *et al.* Gene expression analysis of prostate cancers. *Mol. Carcinog.* **33**, 25–35 (2002).
42. Vanaja, D.K., Cheville, J.C., Iturria, S.J. & Young, C.Y. Transcriptional silencing of zinc finger protein 185 identified by expression profiling is associated with prostate cancer progression. *Cancer Res.* **63**, 3877–3882 (2003).
43. Varambally, S. *et al.* Integrative genomic and proteomic analysis of prostate cancer reveals signatures of metastatic progression. *Cancer Cell* **8**, 393–406 (2005).
44. Wallace, T.A. *et al.* Tumor immunobiological differences in prostate cancer between African-American and European-American men. *Cancer Res.* **68**, 927–936 (2008).
45. Yu, Y.P. *et al.* Gene expression alterations in prostate cancer predicting tumor aggression and preceding development of malignancy. *J. Clin. Oncol.* **22**, 2790–2799 (2004).

ONLINE METHODS

Cell lines. All cell lines were obtained from the American Type Culture Collection. Cell lines were maintained using standard media and conditions. Specifically, VCaP and Du145 cells were maintained in DMEM (Invitrogen) supplemented with 10% FBS and 1% penicillin-streptomycin. LNCaP and 22Rv1 cells were maintained in RPMI 1640 (Invitrogen) supplemented with 10% FBS and 1% penicillin-streptomycin. RWPE cells were maintained in KSF medium (Invitrogen) supplemented with 10 ng/ml epidermal growth factor (EGF; Sigma) and bovine pituitary extract (BPE) and with 1% penicillin-streptomycin. All cell lines were grown at 37 °C in a 5% CO_2 cell culture incubator. All cell lines were genotyped for identity at the University of Michigan Sequencing Core and were tested routinely for Mycoplasma contamination.

Cell lines expressing *SChLAP1* or control constructs were generated by cloning *SChLAP1* or control sequence into the pLenti6 vector (Invitrogen), using pcr8 non-directional Gateway cloning (Invitrogen) as an initial cloning vector, and shuttling to pLenti6 using LR clonase II (Invitrogen) according to the manufacturer's instructions. Stably transfected RWPE and 22Rv1 cells were selected with blasticidin (Invitrogen) for 1 week. For LNCaP and 22Rv1 cells with stable knockdown of *SChLAP1*, cells were transfected with lentiviral constructs encoding *SChLAP1* shRNA or with non-targeting shRNA lentiviral constructs for 48 h. GFP-positive cells were selected with 1 µg/ml puromycin for 72 h. All lentiviruses were generated by the University of Michigan Vector Core.

Tissue samples. Prostate tissues were obtained from the radical prostatectomy series and Rapid Autopsy Program at the University of Michigan tissue core[46]. These programs are part of the University of Michigan Prostate Cancer Specialized Program of Research Excellence (SPORE). All tissue samples were collected with informed consent under an institutional review board (IRB)-approved protocol at the University of Michigan (SPORE in Prostate Cancer (Tissue/Serum/Urine) Bank Institutional Review Board 1994-0481).

RNA isolation and cDNA synthesis. Total RNA was isolated using TRIzol (Invitrogen) and an RNeasy kit (Qiagen) with DNase I digestion according to the manufacturers' instructions. RNA integrity was verified on an Agilent Bioanalyzer 2100 (Agilent Technologies). cDNA was synthesized from total RNA using Superscript III (Invitrogen) and random primers (Invitrogen).

Quantitative RT-PCR. Quantitative RT-PCR was performed using Power SYBR Green MasterMix (Applied Biosystems) on an Applied Biosystems 7900HT Real-Time PCR System. All oligonucleotide primers were obtained from Integrated DNA Technologies (IDT), and primer sequences are listed in **Supplementary Table 7a**. The housekeeping genes *GAPDH*, *HMBS* and *ACTB* were used as loading controls. Fold changes were calculated relative to housekeeping genes and were normalized to the median value in benign samples.

RT-PCR. RT-PCR was performed for primer pairs using Platinum Taq High-Fidelity polymerase (Invitrogen). PCR products were resolved on a 1.0% agarose gel. PCR products were then either sequenced directly (if only a single product was observed) or appropriate gel products were extracted using a Gel Extraction kit (Qiagen) and cloned into pcr4-TOPO vector (Invitrogen). PCR products were bidirectionally sequenced at the University of Michigan Sequencing Core using either gene-specific primers or M13 forward and reverse primers for cloned PCR products. All oligonucleotide primers were obtained from IDT, and primer sequences are listed in **Supplementary Table 7a**.

RACE. 5′ and 3′ RACE were performed using the GeneRacer RLM-RACE kit (Invitrogen) according to the manufacturer's instructions. RACE PCR products were obtained using Platinum Taq High-Fidelity polymerase, the supplied GeneRacer primers and the appropriate gene-specific primers indicated in **Supplementary Table 7a**. RACE PCR products were separated on a 1.5% agarose gel. Gel products were extracted with a Gel Extraction kit, cloned into pcr4-TOPO vectors and sequenced bidirectionally using M13 forward and reverse primers at the University of Michigan Sequencing Core. At least three colonies were sequenced for every RACE PCR product that was gel purified.

siRNA-mediated knockdown. Cells were plated in 100-mm plates at a desired concentration and transfected with 20 µM experimental siRNA oligonucleotides or non-targeting controls twice at 8 h and 24 h after plating. Knockdown was performed with Oligofectamine in OptiMEM medium. Knockdown efficiency was determined by qPCR. siRNA sequences (in sense orientation) for knockdown experiments are listed in **Supplementary Table 7b**. At 72 h after transfection, cells were trypsinized, counted with a Coulter counter and diluted to 1 million cells/ml.

Overexpression. Full-length *SChLAP1* transcript was amplified from LNCaP cells and cloned into the pLenti6 vector along with *LacZ* control sequence. Insert sequences were confirmed by Sanger sequencing at the University of Michigan Sequencing Core. Lentiviruses were generated at the University of Michigan Vector Core. The benign immortalized prostate cell line RWPE was infected with lentiviruses expressing *SChLAP1* or *LacZ*, and stable pools and clones were generated by selection with blasticidin. Similarly, the immortalized cancer cell line 22Rv1 was infected with lentiviruses expressing *SChLAP1* or *LacZ*, and stable pools were generated by selection with blasticidin.

Cell proliferation assays. At 72 h after transfection with siRNA, cells were trypsinized, counted with a Coulter counter and diluted to 1 million cells/ml. For proliferation assays, 10,000 cells were plated in each well of a 24-well plate and grown in regular growth medium. At 48 h and 96 h after plating, cells were collected by trypsinizing and counted using a Coulter counter. All assays were performed in quadruplicate.

Basement membrane matrix invasion assays. For invasion assays, cells were treated with the indicated siRNAs, and, at 72 h after transfection, cells were trypsinized, counted with a Coulter counter and diluted to 1 million cells/ml. Cells were seeded onto basement membrane matrix (EC matrix, Chemicon) present in the insert of a 24-well culture plate. FBS was added to the lower chamber as a chemoattractant. After 48 h, the non-invading cells and EC matrix were gently removed with a cotton swab. Invasive cells located on the lower side of the chamber were stained with crystal violet, air dried and photographed. For colorimetric assays, inserts were treated with 150 µl of 10% acetic acid, and absorbance was measured at 560 nm using a spectrophotometer (GE Healthcare).

shRNA-mediated knockdown. The prostate cancer cell lines LNCaP and 22Rv1 were seeded at 50–60% confluency and were allowed to attach overnight. Cells were transfected with lentiviral constructs expressing *SChLAP1* or non-targeting shRNA as described previously for 48 h. GFP-positive cells were selected with 1 µg/ml puromycin for 72 h. At 48 h after the start of selection, cells were collected for protein and RNA using RIPA buffer or TRIzol, respectively. RNA was processed as described above.

Gene expression profiling. Expression profiling was performed using the Agilent Whole Human Genome Oligo Microarray according to previously published protocols[47]. All samples were run in technical triplicates, comparing knockdown samples treated with *SChLAP1* siRNA to samples treated with non-targeting control siRNA. Expression data were analyzed using the SAM method as described previously[20].

Mouse intracardiac and subcutaneous *in vivo* models. All experimental procedures were approved by the University of Michigan Committee for the Use and Care of Animals (UCUCA).

For the intracardiac injection model, 5×10^5 cells from 1 of 3 experimental cell lines (22Rv1-sh*SChLAP1*-1 or 22Rv1-sh*SChLAP1*-2 (two cell lines expressing *SChLAP1* shRNA) or 22Rv1-shNT (expressing control vector), all with luciferase constructs incorporated) were introduced into CB-17 SCID mice at 6 weeks of age. Female mice were used to minimize endogenous androgen production that might stimulate xenografted prostate cells. We used 15 mice per cell line to ensure adequate statistical power to distinguish phenotypes between groups. Mice used in these studies were randomized by double-blind injection of cell line samples into mice and were monitored for tumor growth by researchers blinded to the study design. Beginning 1 week after injection, bioluminescent imaging of mice was performed weekly using a CCD IVIS system with a 50-mm lens (Xenogen), and the results were analyzed using LivingImage software (Xenogen). When the a mouse reached the determined end point, defined as whole-body region of interest (ROI) of 1 × 10^{10} photons, or became fatally ill, it was euthanized, and the lung and liver were resected. Half of the resected specimen was placed in an immunohistochemistry cassette, incubated in 10% buffered formalin phosphate (Fisher Scientific) for 24 h and transferred to 70% ethanol until further analysis. The other half of each specimen was snap frozen in liquid nitrogen and stored at −80 °C. A specimen was disregarded if the tumor was localized only in the heart. After accounting for these considerations, there were 9 mice analyzed for 22Rv1-shNT cells and 14 mice each analyzed for 22Rv1-sh*SChLAP1*-1 and 22Rv1-sh*SChLAP1*-2 cells.

For the subcutaneous injection model, 1 × 10^6 cells from 1 of the 3 previously described experimental cell lines were introduced into mice (CB-17 SCID), aged 5–7 weeks, with a Matrigel scaffold (BD Matrigel Matrix, BD Biosciences) in the posterior dorsal flank region (n = 10 per cell line). Tumors were measured weekly using a digital caliper, and the end point was defined by tumor volume of 1,000 mm^3. When a mouse reached the end point or became fatally ill, it was euthanized, and the primary tumor was resected. The resected specimen was divided in half: one half was placed in 10% buffer formalin, and the other half was snap frozen. For histological analyses, formalin-fixed, paraffin-embedded mouse livers and lungs were sectioned on a microtome into 5-µm sections on glass slides. Slides were stained with hematoxylin and eosin using standard methods and were analyzed by a board-certified pathologist (L.P.K.).

Immunoblot analysis. Cells were lysed in RIPA lysis buffer (Sigma) supplemented with HALT protease inhibitor (Fisher). Protein blotting analysis was performed with standard protocols using polyvinylidene difluoride (PVDF) membrane (GE Healthcare), and signals were visualized with an enhanced chemiluminescence system as described by the manufacturer (GE Healthcare).

Protein lysates were boiled in sample buffer, and 10 µg of protein was loaded onto an SDS-PAGE gel and run for separation of proteins. Proteins were transferred onto PVDF membrane and blocked for 90 min in blocking buffer (5% milk in a solution of 0.1% Tween-20 in Tris-buffered saline (TBS-T)). Membranes were incubated overnight at 4 °C with primary antibody. After three washes with TBS-T and one wash with TBS, the blot was incubated with HRP-conjugated secondary antibody, and signal was visualized with an enhanced chemiluminescence system as described by the manufacturer. Primary antibodies used included antibody to SNF5 (1:1,000 dilution; Millipore, ABD22, rabbit), SNF5 (1:1,000 dilution; Abcam, ab58209, mouse), β-actin (1:5,000 dilution; Sigma, A5316, mouse) and androgen receptor (1:1,000 dilution; Millipore, 06-680, rabbit).

RIP assays. RIP assays were performed using a Millipore EZ-Magna RIP RNA-Binding Protein Immunoprecipitation kit (Millipore, 17-701) according to the manufacturer's instructions. RIP PCR was performed as qPCR, as described above, using total RNA as input controls. We used 1/150 volume of the RIP RNA product per PCR reaction. Antibodies used for RIP included rabbit polyclonal IgG (Millipore, PP64) and antibodies to SNRNP70 (Millipore, CS203216), SNF5 (Millipore, ABD22, rabbit), SNF5 (Abcam, ab58209, mouse) and androgen receptor (Millipore, 06-680, rabbit), and 5–7 µg of antibody was used per RIP reaction. All RIP assays were performed in biological duplicate. For UV-crosslinked RIP experiments, cells were subjected to 400 J of 254 nM UV light twice and were then collected for RIP experiments as described above.

ChIP assays. ChIP assays were performed as described previously[11,12] using antibody for SNF5 (Millipore, ABD22, rabbit) and rabbit IgG (Millipore, PP64B). Briefly, approximately 1 million cells were cross-linked per antibody for 10–15 min with 1% formaldehyde, and crosslinking was inactivated by incubation with 0.125 M glycine for 5 min at room temperature. Cells were rinsed with cold PBS three times, and cell pellets were resuspended in lysis buffer supplemented with protease inhibitors. Chromatin was sonicated to an average length of 500 bp and centrifuged to remove debris, and supernatants containing chromatin fragments were incubated with protein A or protein G beads to reduce non-specific binding. Beads were then removed, and supernatants were incubated with 6 µg of antibody overnight at 4 °C. Fresh beads were added and incubated with protein-chromatin-antibody complexes for 2 h at 4 °C, washed twice with 1× dialysis buffer and four times with IP wash buffer, and eluted in 150 µl of IP elution buffer[12]. One-tenth of the ChIP reaction was taken for protein evaluation for validation of pulldown. Cross-linking was reversed by incubating eluted products with 0.3 M NaCl at 65 °C overnight. ChIP products were cleaned with the USB PrepEase kit. ChIP experiments were validated for specificity of the antibody by protein blotting.

ChIP-seq experiments. Paired-end ChIP-seq libraries were generated following the Illumina ChIP-seq protocol with minor modifications. DNA isolated by ChIP assay was subjected to end repair and A tailing before ligation with Illumina adaptors. Samples were purified using AMPure beads (Beckman Coulter) and PCR enriched with a combination of specific index primers and PE2.0 primer under the following conditions: 98 °C (30 s), 65 °C (30 s) and 72 °C (40 s, with the addition of 4 s per cycle). After 14 cycles of amplification, a final extension at 72 °C for 5 min was carried out. Barcoded libraries were size selected using 3% NuSieve Agarose gels (Lonza) and subjected to an additional PCR enrichment step. Libraries were analyzed and quantified using a Bioanalyzer instrument (Agilent Technologies) before they were subjected to paired-end sequencing using the Illumina HiSeq platform.

CAM assays. CAM assays were performed as previously described[19]. Briefly, fertilized chicken eggs were incubated in a rotary humidified incubator at 38 °C for 10 d. CAM was released by applying a mild amount of pressure to the hole over the air sac and cutting a 1-cm^2 window encompassing a second hole near the allantoic vein. Approximately 2 million cells in 50 µl of medium were implanted in each egg, windows were sealed, and eggs were returned to a stationary incubator.

For local invasion and intravasation experiments, the upper and lower CAMs were isolated after 72 h. Upper CAMs were processed and stained for chicken collagen IV (immunofluorescence) or human cytokeratin (immunohistochemistry) as previously described[19].

For metastasis assays, embryonic livers were isolated on day 18 of embryonic growth and analyzed for the presence of tumor cells by quantitative human Alu-specific PCR. Genomic DNA isolates from lower CAMs and livers were prepared using the Puregene DNA purification system (Qiagen), and quantification with human Alu-specific PCR was performed as described[19]. Fluorogenic TaqMan qPCR probes were generated as described above and used to determine DNA copy number.

For xenograft growth assays with RWPE cells, embryos were sacrificed on day 18, and extraembryonic xenografts were excised and weighed.

***In situ* hybridization.** *In situ* hybridization assays were performed as a commercial service from Advanced Cell Diagnostics, Inc. Briefly, cells in the clinical specimens were fixed and permeablized using xylene, ethanol and protease to allow for probe access. Slides were boiled in pretreatment buffer for 15 min and rinsed in water. Next, two independent target probes were hybridized to *SChLAP1* RNA at 40 °C for 2 h, with this pair of probes creating a binding site for a preamplifier. After this incubation, the preamplifier was hybridized to the target probes at 30 °C and amplified with six cycles of hybridization followed by two washes. Cells were counterstained to visualize signal. Finally, slides were stained with hematoxylin and eosin, dehydrated with 100% ethanol and xylene and mounted in a xylene-based mounting medium.

***In vitro* translation.** Full-length *SChLAP1*, *PCAT-1* or *GUS* positive control sequences were cloned into the PCR2.1 entry vector (Invitrogen). Insert sequences were confirmed by Sanger sequencing at the University of Michigan Sequencing Core. *In vitro* translation assays were performed with the TnT Quick Coupled Transcription/Translation System (Promega) with 1 mM methionine and Transcend Biotin-Lysyl-tRNA (Promega) according to the manufacturer's instructions.

ChIRP assays. ChIRP assays were performed as previously described[48]. Briefly, antisense DNA probes targeting the full-length *SChLAP1* sequence

doi:10.1038/ng.2771

were designed using the online designer at Stellaris (see URLs). Fifteen probes spanning the entire transcript and unique to the *SChLAP1* sequence were chosen. Additionally, ten probes were designed against *TERC* RNA as a positive control, and 24 probes were designed against *LacZ* RNA as a negative control. All probes were synthesized with 3′ biotinylation (IDT). Sequences of all probes are listed in **Supplementary Table 8**. RWPE cells overexpressing *SChLAP1* isoform 1 were grown to 80% confluency in 100-mm cell culture dishes. Two dishes were used for each probe set. Before being collected, cells were rinsed with 1× PBS and cross-linked with 1% glutaraldehyde (Sigma) for 10 min at room temperature. Cross-linking was quenched by incubation with 0.125 M glycine for 5 min at room temperature. Cells were rinsed twice with 1× PBS, collected and pelleted at 1,500*g* for 5 min. Nuclei were isolated using the Pierce NE-PER Nuclear Protein Extraction kit. Nuclear pellets were resuspended in 100 mg/ml cell lysis buffer (50 mM Tris, pH 7.0, 10 mM EDTA, 1% SDS, and, added before use, 1 mM dithiothreitol (DTT), phenylmethylsulphonyl fluoride (PMSF), protease inhibitor and Superase-In (Invitrogen)). Lysates were placed on ice for 10 min and sonicated using a Bioruptor (Diagenode) at the highest setting with 30-s on and 45-s off cycles until lysates were completely solubilized. Cell lysates were diluted in twice the volume of hybridization buffer (500 mM NaCl, 1% SDS, 100 mM Tris, pH 7.0, 10 mM EDTA, 15% formamide, and, added before use, DTT, PMSF, protease inhibitor and Superase-In), and 100 nM probes were added to the diluted lysates. Hybridization was carried out by end-over-end rotation at 37 °C for 4 h. Magnetic streptavidin C1 beads were prepared by washing three times in cell lysis buffer and were then added to each hybridization reaction at a concentration of 100 µl per 100 pmol of probe. Reactions were incubated at 37 °C for 30 min with end-over-end rotation. Bead-probe-RNA complexes were captured with magnetic racks (Millipore) and washed five times with 1 ml wash buffer (2× SSC, 0.5% SDS, fresh PMSF added). After the final wash, 20% of the sample was used for RNA isolation, and 80% of the sample was used for protein isolation. For RNA elution, beads were resuspended in 200 µl of RNA proteinase K buffer (100 mM NaCl, 10 mM Tris, pH 7.0, 1 mM EDTA, 0.5% SDS) and 1 mg/ml proteinase K (Ambion). Samples were incubated at 50 °C for 45 min and then boiled for 10 min. RNA was isolated using 500 µl of TRIzol reagent and the miRNeasy kit (Qiagen) with on-column DNase digestion (Qiagen). RNA was eluted with 10 µl of water and then analyzed by quantitative RT-PCR for the detection of enriched transcripts. For protein elution, beads were resuspended in three times the original volume of DNase buffer (100 mM NaCl, 0.1% NP-40), and protein was eluted with a cocktail of 100 µg/ml RNase A (Sigma-Aldrich), 0.1 U/ml RNase H (Epicenter) and 100 U/ml DNase I (Invitrogen) at 37 °C for 30 min. Eluted protein samples were supplemented with NuPAGE LDS Sample Buffer (Novex) and NuPAGE Sample Reducing Agent (Novex) to a final concentration of 1× each and then boiled for 10 min before SDS-PAGE protein blot analysis using an antibody to SNF5 (Millipore).

RNA-seq library preparation. Total RNA was extracted from healthy and cancer cell lines and subject tissues, and RNA quality was assessed via Agilent Bioanalyzer. Transcriptome libraries from the mRNA fractions were generated following the RNA-seq protocol (Illumina). Each sample was sequenced in a single lane with the Illumina Genome Analyzer II (with a 40- to 80-nt read length) or with the Illumina HiSeq 2000 (with a 100-nt read length) according to published protocols[11,49]. For strand-specific library construction, we employed the dUTP method of second-strand marking as described previously[50].

Statistical analyses for experimental studies. All data are presented as means ± s.e.m. All experimental assays were performed in duplicate or triplicate. Statistical analyses shown in figures represent Fisher's exact tests or two-tailed *t* tests, as indicated. For details regarding the statistical methods employed during microarray, RNA-seq and ChIP-seq data analysis, see below.

Nomination of *SChLAP1* as an outlier using RNA-seq data. We nominated *SChLAP1* as a prostate cancer outlier as described[11]. Briefly, a modified COPA analysis was performed on the 81 tissue samples in the cohort. Reads per kilobase per million mapped reads (RPKM) expression values were used and shifted by 1.0 to avoid division by zero. COPA analysis included the following steps: (i) gene expression values were median centered, using the median expression value for the gene across all samples in the cohort, which sets the gene's median to zero; (ii) the median absolute deviation (MAD) was calculated for each gene, and each gene expression value was then scaled by its MAD; (iii) the 80th, 85th, 90th and 98th percentiles of the transformed expression values were calculated for each gene, the average of those four values was taken, and genes were then ranked according to this 'average percentile', which generated a list of outlier genes arranged by importance; and (iv) finally, genes showing an outlier profile in the benign samples were discarded.

LNCaP ChIP-seq data. Sequencing data from GSE14097 were downloaded from GEO. Reads from the LNCAP H3K4me3 and H3K36me3 ChIP-seq samples were mapped to human genome version hg19 using BWA 0.5.9 (ref. 51). Peak calling was performed using MACS[52] according to published protocols[53]. Data were visualized using the UCSC Genome Browser[54].

RWPE ChIP-seq data. Sequencing data from RWPE SNF5 ChIP-seq samples were mapped to human genome version hg19 using the BWA 0.5.9 algorithm[51]. Although we performed paired-end sequencing, the ChIP-seq reads were processed as single-end reads to adhere to our preexisting analysis protocol. Basic read alignment statistics are listed in **Supplementary Table 6a**. Peak calling was performed with respect to an IgG control using the MACS algorithm[52]. We bypassed the model-building step of MACS (using the '–nomodel' flag) and specified a shift size equal to half the library fragment size determined by the Agilent Bioanalyzer (using the '–shiftsize' option). For each sample, we ran the CEAS program and generated genome-wide reports[55]. We retained peaks with an FDR less than 5% (peak calling statistics across multiple FDR thresholds are shown in **Supplementary Table 6b**). We then aggregated SNF5 peaks from the RWPE-*LacZ*, RWPE–*SChLAP1* isoform 1 and RWPE–*SChLAP1* isoform 2 samples using the 'union' of the genomic peak intervals. We intersected peaks with RefSeq protein-coding genes and found that 1,299 peaks occurred within 1 kb of TSSs. We counted the number of reads overlapping each of these promoter peaks across each sample using a custom Python script and used the DESeq R package[56] version 1.6.1 to compute the normalized fold change between RWPE-*LacZ* and RWPE-*SChLAP1* (both isoforms). We observed that 389 of the 1,299 promoter peaks had at least a 2-fold average decrease in SNF5 binding. This set of 389 genes was subsequently used as a gene set for GSEA (**Supplementary Table 6c**).

Microarray experiments. We performed two-color microarray gene expression profiling of 22Rv1 and LNCaP cells treated with two independent siRNAs targeting *SChLAP1* as well as control non-targeting siRNAs. These profiling experiments were run in technical triplicate for a total of 12 arrays (6 from 22Rv1 and 6 from LNCaP). Additionally, we profiled 22Rv1 and LNCaP cells treated with independent siRNAs targeting SWI/SNF component *SMARCB1* as well as control non-targeting siRNAs. These profiling experiments were run as biological duplicates for a total of four arrays (two cell lines × two independent siRNAs × one protein). Finally, we profiled RWPE cells expressing two different *SChLAP1* isoforms as well as the control *LacZ* gene. These profiling experiments were run in technical duplicate for a total of four arrays (two from RWPE–*SChLAP1* isoform 1 and two from RWPE–*SChLAP1* isoform 2).

Processing to determine ranked gene expression lists. All of the microarray data were represented as $\log_2$ fold change between targeting versus control siRNAs. We used the CollapseDataset tool provided by the GSEA package to convert Agilent Probe IDs to gene symbols. Genes whose expression was measured by multiple probes were consolidated using the median values obtained with these probes. We then ran one-class SAM analysis from the Multi-Experiment Viewer application and ranked all genes by the difference between observed versus expected statistics. These ranked gene lists were imported to GSEA version 2.07.

***SChLAP1* siRNA knockdown microarrays.** For the 22Rv1 and LNCaP *SChLAP1* knockdown experiments, we ran the GseaPreRanked tool to discover enriched gene sets in MSigDB[22] version 3.0. Lists of positively and negatively enriched concepts were interpreted manually.

doi:10.1038/ng.2771

***SMARCB1* siRNA knockdown microarrays.** For each *SMARCB1* knockdown experiment, we nominated genes that were altered by an average of at least twofold. These signatures of putative SNF5 target genes were then used to assess enrichment of *SChLAP1*-regulated genes using the GseaPreRanked tool. Additionally, we nominated genes whose expression changed by an average of twofold or greater across *SMARCB1* knockdown experiments and quantified the enrichment for *SChLAP1* target genes using GSEA.

RWPE *SChLAP1* expression microarrays. RWPE-*SChLAP1* versus RWPE-*LacZ* expression profiles were ranked using SAM analysis as described above. A total of 1,245 genes were significantly over- or underexpressed and are shown in **Supplementary Table 6d**. A q value of 0.0 in this SAM analysis signifies that no permutation generated a more significant difference between observed and expected gene expression ratios. The ranked gene expression list was used as input for the GseaPreRanked tool and compared against SNF5 ChIP-seq promoter peaks that decreased by >2-fold in RWPE cells overexpressing *SChLAP1*. Of the 389 genes in the ChIP-seq gene set, 250 were profiled by the Agilent HumanGenome microarray chip and were present in the GSEA gene symbol database. The expression profile across these 250 genes is shown in **Supplementary Table 6e**.

RNA-seq data. We assembled an RNA-seq cohort from prostate cancer tissues sequenced at multiple institutions. We included data from 12 primary tumors and 5 benign tissues published in GEO (GSE22260)[57], from 16 primary tumors and 3 benign tissues released in the database of Genotypes and Phenotypes (dbGAP) (phs000310.v1.p1)[58] and from 17 benign, 57 primary and 14 metastatic tumors sequenced by our own institution and released in dbGAP (phs000443.v1.p1). Sample information is shown in **Supplementary Table 1a**, and sequencing library information is shown in **Supplementary Table 1b**.

RNA-seq alignment and gene expression quantification. Sequencing data were aligned using TopHat[59] version 1.3.1 against the Ensembl GRCh37 human genome build. Known introns (Ensembl release 63) were provided to TopHat. Gene expression across genes in Ensembl version 63 and the *SChLAP1* transcript was quantified by HT-Seq version 0.5.3p3 using the script 'htseq-count'. Reads were counted without respect to strand to avoid bias between unstranded and strand-specific library preparation methods. This bias results from the inability to resolve reads in regions where two genes on opposite strands overlap in the genome.

RNA-seq differential expression analysis. Differential expression analysis was performed using R package DESeq[56] version 1.6.1. Read counts were normalized using the 'estimateSizeFactors' function, and variance was modeled by the 'estimateDispersions' function. Statistics on differential expression were computed by the 'nbinomTest' function. We called differentially expressed genes by imposing adjusted P-value cutoffs for cancer versus benign samples ($P_{\mathrm{adj}} < 0.05$), metastasis versus primary samples ($P_{\mathrm{adj}} < 0.05$) and Gleason score of 8+ versus 6 ($P_{\mathrm{adj}} < 0.10$). Heatmap visualizations of these analyses are presented as **Supplementary Figure 5**.

RNA-seq correlation analysis. Read count data were normalized using functions from the R package DESeq version 1.6.1. Adjustments for library size were made using the 'estimateSizeFactors' function, and variance was modeled using the 'estimateDispersions' function using the parameters 'method=blind' and 'sharingMode=fit-only'. Next, raw read count data were converted to pseudocounts using the 'getVarianceStabilizedData' function. Gene expression levels were then mean centered and standardized using the 'scale' function in R. Pearson's correlation coefficients were computed between each gene of interest and all other genes. Statistical significance of Pearson's correlations was determined by comparison to correlation coefficients achieved with 1,000 random permutations of the expression data. We controlled for multiple-hypothesis testing using the 'qvalue' package in R. The *SChLAP1* correlation signature of 253 genes was determined by imposing a cutoff of $q < 0.05$.

Oncomine concepts analysis of the *SChLAP1* signature. We separated the 253 genes with expression levels significantly correlated with *SChLAP1* into positively and negatively correlated gene lists. We imported these gene lists into Oncomine as custom concepts. We then nominated significantly associated prostate cancer concepts with odds ratio > 3.0 and $P < 1 \times 10^{-6}$. We exported these results as the nodes and edges of a concept association network and visualized the network using Cytoscape version 2.8.2. Node positions were computed using the Force-Directed Layout algorithm in Cytoscape using the odds ratio as the edge weight. Node positions were subtly altered manually to enable better visualization of node labels.

Association of correlation signatures with Oncomine concepts. We applied our RNA-seq correlation analysis procedure to the genes *SChLAP1*, *EZH2*, *PCA3*, *AMACR* and *ACTB*. For each gene, we created signatures from the top 5% of positively and negatively correlated genes (**Supplementary Table 3**). We performed a large meta-analysis of these correlation signatures across Oncomine data sets corresponding to disease outcome (Glinsky Prostate and Setlur Prostate), metastatic disease (Holzbeierlein Prostate, Lapointe Prostate, LaTulippe Prostate, Taylor Prostate 3, Vanaja Prostate, Varambally Prostate and Yu Prostate), advanced Gleason score (Bittner Prostate, Glinsky Prostate, Lapointe Prostate, LaTulippe Prostate, Setlur Prostate, Taylor Prostate 3 and Yu Prostate) and localized cancer (Arredouani Prostate, Holzbeierlein Prostate, Lapointe Prostate, LaTulippe Prostate, Taylor Prostate 3, Varambally Prostate and Yu Prostate). We also incorporated our own concept signatures for metastasis, advanced Gleason score and localized cancer determined from our RNA-seq data. For each concept, we downloaded the gene signatures corresponding to the top 5% of genes up- and downregulated. Pairwise signature comparisons were performed using a one-sided Fisher's exact test. We controlled for multiple-hypothesis testing using the 'qvalue' package in R. We considered concept pairs with $q < 0.01$ and odds ratio > 2.0 as significant. In cases where a gene signature associated with both the over- and underexpression gene sets from a single concept, only the most significant result (as determined by odds ratio) is shown.

Analysis of *SChLAP1* and *SMARCB1* expression signatures. Gene signatures obtained with knockdown of *SChLAP1* and *SMARCB1* were generated from Agilent gene expression microarray data sets. For each cell line, we obtained a single vector of per-gene fold changes by averaging technical replicates and then taking the median across biological replicates. We merged the results from individual cell line using the median of the changes in 22Rv1 and LNCaP cells. Venn diagram plots were produced using the BioVenn website[60]. We then compared the top 10% of upregulated and downregulated genes with knockdown of *SChLAP1* and *SMARCB1* to gene signatures downloaded from the Taylor Prostate 3 data set in the Oncomine database. We performed signature comparison using one-sided Fisher's exact tests and controlled for multiple testing using the R package 'qvalue'. Signature comparisons with $q < 0.05$ were considered significantly enriched. We plotted the odds ratios from significant comparisons using the 'heatmap.2' function in the 'gplots' R package.

Kaplan-Meier survival analysis based on the *SChLAP1* gene signature. We downloaded prostate cancer expression profiling data and clinical annotations from GSE8402, published by Setlur *et al.*[17]. We intersected the 253-gene *SChLAP1* signature with the genes in this data set and found 80 genes in common. We then assigned *SChLAP1* expression scores to each patient sample in the cohort using the unweighted sum of standardized expression levels across the 80 genes. Given that we observed *SChLAP1* expression in approximately 20% of prostate cancer samples, we used the 80th percentile of *SChLAP1* expression scores as the threshold for 'high' versus 'low' scores. We then performed 10-year survival analysis using the 'survival' package in R and computed statistical significance using the log-rank test.

Additionally, we imported the 253-gene *SChLAP1* signature into Oncomine to download the expression data for 167 of the 253 genes profiled by the Glinsky prostate data set[16]. We assigned *SChLAP1* expression scores in a similar fashion and designated the top 20% of patients as having 'high' *SChLAP1* scores. We performed survival analysis using the time to biochemical prostate-specific antigen (PSA) recurrence and computed statistical significance as described above.

PhyloCSF analysis. We obtained 46-way multi-alignment FASTA files for *SChLAP1*, *HOTAIR*, *GAPDH* and *ACTB* using the 'Stitch Gene blocks' tool within the Galaxy bioinformatics framework. We evaluated each gene for the likelihood that it represented a protein-coding region using PhyloCSF software (version released 28 October 2012). Each gene was evaluated using the phylogeny from 29 mammals (available by default within PhyloCSF) in any of the 3 reading frames. Scores are measured in decibans and represent the likelihood ratio that a sequence is protein-coding rather than noncoding.

Mayo Clinic cohort analyses. Subjects were selected from a cohort of individuals from the Mayo Clinic with high-risk prostate cancer who had undergone radical prostatectomy. The cohort was defined as 1,010 men with high-risk prostate cancer who underwent radical prostatectomy between 2000 and 2006, of whom 73 developed clinical progression (defined as individuals with systemic disease as evidenced by positive bone or computed tomography (CT) scan)[61]. High risk of recurrence was defined by preoperative PSA levels of >20 ng/ml, pathological Gleason score of 8–10, seminal vesicle invasion (SVI) or Gleason, PSA, seminal vesicle and margin (GPSM) score of ≥10 (ref. 62). The subcohort incorporated all 73 subjects with clinical progression to systemic disease and a random sampling of 20% of the entire cohort (202 men, including 19 with clinical progression). The total case-cohort study included 256 subjects, and tissue specimens were available from 235 subjects. The subcohort was previously used to validate a genomic classifier for predicting clinical progression[61].

Tissue preparation. Formalin-fixed, paraffin-embedded samples of human prostate adenocarcinoma prostatectomies were collected from subjects with informed consent at the Mayo Clinic according to an IRB-approved protocol. Pathological review of tissue sections stained with hematoxylin and eosin was used to guide macrodissection of the tumor from surrounding stromal tissue in three to four 10-µm sections. The index lesion was considered as the dominant lesion by size.

RNA extraction and microarray hybridization. For the validation cohort, total RNA was extracted and purified using a modified protocol for the commercially available RNeasy FFPE nucleic acid extraction kit (Qiagen). RNA concentrations were calculated using a Nanodrop ND-1000 spectrophotometer (Nanodrop Technologies). Purified total RNA was subjected to whole-transcriptome amplification using the WT-Ovation FFPE system according to the manufacturer's recommendation with minor modifications (NuGen). For the validation, only the Ovation FFPE WTA System was used. Amplified products were fragmented and labeled using the Encore Biotin Module (NuGen) and hybridized to Affymetrix Human Exon (HuEx) 1.0 ST GeneChips following the manufacturer's recommendations.

Microarray expression analysis. Normalization and summarization of the microarray samples was performed with the frozen Robust Multiarray Average (fRMA) algorithm using custom frozen vectors. These custom vectors were created using the vector creation methods described previously[63]. Quantile normalization and robust weighted average methods were used for normalization and summarization, respectively, as implemented in fRMA.

Statistical analysis. Given the exon-intron structure of isoform 1 of *SChLAP1*, all probe selection regions (or PSRs) that fell within the genomic span of *SChLAP1* were inspected for overlap with any of the exons of this gene. One PSR, 2518129, was found to be fully nested within exon 3 of *SChLAP1* and was used for further analysis as a representative PSR for this gene. The PAM (Partition Around Medoids) unsupervised clustering method was used on the expression values of all clinical samples to define two groups with high and low expression of *SChLAP1*.

Statistical analysis on the association of *SChLAP1* with clinical outcomes was carried out using three endpoints: (i) biochemical recurrence, defined as two consecutive increases in serum PSA of ≥0.2 ng/ml after radical prostatectomy; (ii) clinical progression, defined as a positive CT or bone scan; and (iii) prostate cancer–specific mortality.

For the clinical progression end point, all subjects with clinical progression were included in the survival analysis, whereas controls in the subcohort were weighted in a fivefold manner to be representative of individuals from the original cohort. For the prostate cancer–specific mortality end point, cases who did not die from prostate cancer were omitted, and weighting was applied in a similar manner. For biochemical recurrence, because the case cohort was designed on the basis of the clinical progression end point, resampling of subjects with biochemical recurrence and the subcohort was performed to have a representative of the selected individuals with biochemical recurrence from the original cohort.

46. Rubin, M.A. *et al.* Rapid ("warm") autopsy study for procurement of metastatic prostate cancer. *Clin. Cancer Res.* **6**, 1038–1045 (2000).
47. Tomlins, S.A. *et al.* Role of the *TMPRSS2-ERG* gene fusion in prostate cancer. *Neoplasia* **10**, 177–188 (2008).
48. Chu, C., Qu, K., Zhong, F.L., Artandi, S.E. & Chang, H.Y. Genomic maps of long noncoding RNA occupancy reveal principles of RNA-chromatin interactions. *Mol. Cell* **44**, 667–678 (2011).
49. Maher, C.A. *et al.* Chimeric transcript discovery by paired-end transcriptome sequencing. *Proc. Natl. Acad. Sci. USA* **106**, 12353–12358 (2009).
50. Levin, J.Z. *et al.* Comprehensive comparative analysis of strand-specific RNA sequencing methods. *Nat. Methods* **7**, 709–715 (2010).
51. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
52. Zhang, Y. *et al.* Model-based analysis of ChIP-Seq (MACS). *Genome Biol.* **9**, R137 (2008).
53. Feng, J., Liu, T. & Zhang, Y. Using MACS to identify peaks from ChIP-Seq data. *Curr. Protoc. Bioinformatics* **Chapter 2** Unit 2.14 (2011).
54. Kent, W.J. *et al.* The human genome browser at UCSC. *Genome Res.* **12**, 996–1006 (2002).
55. Shin, H., Liu, T., Manrai, A.K. & Liu, X.S. CEAS: *cis*-regulatory element annotation system. *Bioinformatics* **25**, 2605–2606 (2009).
56. Anders, S. & Huber, W. Differential expression analysis for sequence count data. *Genome Biol.* **11**, R106 (2010).
57. Kannan, K. *et al.* Recurrent chimeric RNAs enriched in human prostate cancer identified by deep sequencing. *Proc. Natl. Acad. Sci. USA* **108**, 9172–9177 (2011).
58. Pflueger, D. *et al.* Discovery of non-*ETS* gene fusions in human prostate cancer using next-generation RNA sequencing. *Genome Res.* **21**, 56–67 (2011).
59. Trapnell, C., Pachter, L. & Salzberg, S.L. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* **25**, 1105–1111 (2009).
60. Hulsen, T., de Vlieg, J. & Alkema, W. BioVenn—a web application for the comparison and visualization of biological lists using area-proportional Venn diagrams. *BMC Genomics* **9**, 488 (2008).
61. Karnes, R.J. *et al.* Validation of a genomic classifier that predicts metastasis following radical prostatectomy in an at risk patient population. *J. Urol.* doi:10.1016/j.juro.2013.06.017 (11 June 2013).
62. Blute, M.L., Bergstralh, E.J., Iocca, A., Scherer, B. & Zincke, H. Use of Gleason score, prostate specific antigen, seminal vesicle and margin status to predict biochemical failure after radical prostatectomy. *J. Urol.* **165**, 119–125 (2001).
63. Vergara, I.A. *et al.* Genomic "dark matter" in prostate cancer: exploring the clinical utility of ncRNA as biomarkers. *Front. Genet.* **3**, 23 (2012).